PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Early Processing of Visual Information

D. Marr

Phil. Trans. R. Soc. Lond. B 1976 **275**, 483-519
doi: 10.1098/rstb.1976.0090

EARLY PROCESSING OF VISUAL INFORMATION

By D. MARR
The Artificial Intelligence Laboratory, Massachusetts Institute of Technology,
545, *Technology Square, Cambridge, Mass.* 02139, *U.S.A.*

(*Communicated by S. Brenner, F.R.S. – Received* 23 *January* 1975 – *Revised* 23 *February* 1976)

[Plates 1 and 2]

CONTENTS

An introduction is given to a theory of early visual information processing. The theory has been implemented, and examples are given of images at various stages of analysis.

It is argued that the first step of consequence is to compute a primitive but rich description of the grey-level changes present in an image. The description is expressed in a vocabulary of kinds of intensity change (EDGE, SHADING-EDGE, EXTENDED-EDGE, LINE, BLOB etc.). Modifying parameters are bound to the elements in the description, specifying their POSITION, ORIENTATION, TERMINATION points, CONTRAST, SIZE and FUZZINESS. This description is obtained from the intensity array by fixed techniques, and it is called the *primal sketch*.

For most images, the primal sketch is large and unwieldy. The second important step in visual information processing is to group its contents in a way that is appropriate for later recognition.

From our ability to interpret drawings with little semantic content, one may infer the presence in our perceptual equipment of symbolic processes that can define 'place-tokens' in an image in various ways, and can group them according to certain rules. Homomorphic techniques fail to account for many of these grouping phenomena, whose explanations require mechanisms of construction rather than mechanisms of detection.

The necessary grouping of elements in the primal sketch may be achieved by a mechanism that has available the processes inferred from above, together with the ability to select items by first order discriminations acting on the elements' parameters. Only occasionally do these mechanisms use downward-flowing information about the contents of the particular image being processed.

It is argued that 'non-attentive' vision is in practice implemented by these grouping operations and first order discriminations acting on the primal sketch. The class of computations so obtained differs slightly from the class of second order operations on the intensity array.

The extraction of a form from the primal sketch using these techniques amounts to the separation of figure from ground. It is concluded that most of the separation can be carried out by using techniques that do not depend upon the particular image in question. Therefore, figure-ground separation can normally *precede* the description of the shape of the extracted form.

Up to this point, higher-level knowledge and purpose are brought to bear on only a few of the decisions taken during the processing. This relegates the widespread use of downward-flowing information to a later stage than is found in current machine-vision programs, and implies that such knowledge should influence the control of, rather than interfering with, the actual data-processing that is taking place lower down.

Introduction

The vision problem begins with a large grey-level intensity array, and culminates in a description that depends on that array, and on the purpose for which it is being viewed. The question of interest is what has to go on in between. This article outlines the first part of a theory of visual information processing, and covers the analysis up to about the level of figure-ground separation. The theory is restricted to single frame, monochromatic, monocular images without specularities, reflexions, translucency, transparency or light sources. It is argued that the first step of consequence is to compute a primitive but rich description of the grey-level changes present in an image, and that all subsequent computations are implemented as manipulations of that description. The description itself is called the *primal sketch*. The processes that compute it, and most of the processes that operate directly on it, do not significantly depend upon the particular contents of the image.

The approach taken here rests upon the observation that a drawing of a scene adequately represents the scene, despite the very different grey-level image to which it gives rise. It therefore seems reasonable to suppose that the artist's local symbols are in correspondence with natural symbols, that are computed out of the image during the normal course of its interpretation.

The idea that visual processing should commence with the extraction of a more or less elaborate line-drawing is not a new one, but its successful implementation has proved elusive. Several edge-detection algorithms have been proposed (Hueckel 1971, 1973; McCleod 1970; Rosenfeld & Thurston 1971; Rosenfeld, Thurston & Lee 1972; Horn 1973), but as their proliferation suggests, the results of applying them to natural images have proved generally unsatisfactory. This has led some to believe that an adequate line-drawing of a scene cannot be computed unless hypotheses about what is present are allowed to influence quite early stages in the processing (Shirai 1973; Freuder 1975).

How much independent pre-processing can usefully be carried out? Do the different stages in recognition have to interact in a rich and complex way, or may they be implemented in modules that are to a first approximation independent? These questions do not depend upon the particular hardware (wet or dry) in which the processing is implemented. We need to answer them before we can address 'higher-level' problems, because the nature of the answers determines the overall strategy that subsequent processes must employ.

General principles

Several lessons have been learnt over the last ten years from the experience of designing and implementing large symbolic computer programs. These lessons may be expressed as four principles for the organization of complex symbolic processes. Because of the need to refer to them, and because recognition and other advanced biological computations are complex symbolic processes, I attempt to set out these principles here.

Principle of explicit naming

Whenever a collection of data is to be described, discussed or manipulated as a whole, it should first be given a *name*. This forms the data into an entity in its own right, permits properties to be assigned to it, and allows other structures and processes to refer to it. The act of naming is the distinguishing mark of symbolic computation, and this insight was the single most important idea behind the invention of the programming language LISP (McCarthy *et al.* 1963).

Principle of modular design

Any large computation should be split up and implemented as a collection of small sub-parts that are as nearly independent of one another as the overall task allows. If a process is not designed in this way, a small change in one place will have consequences in many other places. This means that the process as a whole becomes extremely difficult to debug or to improve, whether by a human designer or in the course of natural evolution, because a small change to improve one part has to be accompanied by many simultaneous compensating changes elsewhere.

Principle of least commitment

The principle of least commitment states that one should never do something that may later have to be undone, and I believe that it applies to all situations in which performance is fluent. It is frequently the case during the execution of a recognition task that there are a number of possible interpretations of a particular datum, but that there is not yet sufficient evidence to decide between them. In such cases, one should never become committed to one of the possibilities prematurely, because of the damage that knowledge associated with that possibility and not with the others can subsequently do.

There are two escapes from situations in which the principle is about to be violated. One is to 'wait and see', hopeful that the rival possibilities can be maintained without causing memory overflow until information becomes available that can select the correct interpretation. Marcus (1974) has conjectured that the structure of English syntax is such that a wait-and-see parser never has to wait very long before seeing. The other escape is to restructure the problem, by breaking the computation into more steps, by increasing the vocabulary for expressing the possible choices, and by adding more diagnostics for deciding between rival possibilities. The sheer volume of information rules out a wait-and-see approach to early visual processing, so only the second alternative is a real option there. My experience has been that if one has to disobey the principle of least commitment, one is either doing something wrong, or something very difficult.

An application of the principle is frequently accompanied by a particular style of computation called *constraint analysis*, or *filtering*. We shall meet it later in this article. Where several possibilities compete for the privilege of describing a particular datum, there usually exist constraints or measures of preference that operate among them. The act of filtering the possibilities using the constraints is a distinctive style of computation, somewhat reminiscent of relaxation techniques for solving complex problems in structural engineering. Constraint analysis was first used effectively in a vision program by Waltz (1975). A neural implementation of essentially this technique was given by Marr (1971, § 3.1.2).

Principle of graceful degradation

The final principle is designed to ensure that wherever possible, degrading the data will not prevent one from delivering at least some of the answer. It amounts to a condition on the continuity of the relation between descriptions computed at different stages in the processing. For example, it would be foolish not to require that a 'rough' two dimensional description, of the kind that a vision system might compute out of a drawing, should enable it to compute a 'rough' three dimensional description of what the drawing represents.

Early processing: computing the primal sketch

The primal sketch consists of a primitive but rich description of the intensity changes that are present in an image. This description consists of a set of assertions, expressed in terms of a vocabulary of symbols and modifiers that are powerful enough to capture all of the important information in an intensity array. An example of such an assertion might be.

```
(SHADING-EDGE (POSITION (34 48) (73 48))
              (CONTRAST 34)
              (FUZZINESS 17)
              (ORIENTATION 0))
```

The design of a method for achieving this rests on two primary decisions; what types of intensity change are to be detected, and how expressive is the vocabulary in terms of which these changes are to be described?

One dimensional intensity profiles

In an empirical study, Herskovitz & Binford (1970, pp. 19, 53, 55) found that the most common intensity changes in images of scenes composed of polyhedral objects were step changes, bumps, and roof-shaped profiles. Our experience adds some others for more general scenes (see figure 2). The detection of roof-shaped intensity changes requires a sensitivity to changes in intensity gradient. The human visual system has long been known to be sensitive to such changes (Mach bands, Ratliff 1965), but of the edge-detection algorithms referred to in the introduction, only the Binford–Horn line-finder (Horn 1973) incorporates a sensitivity to the second derivative of intensity. There is no evidence that humans are sensitive to higher derivatives.

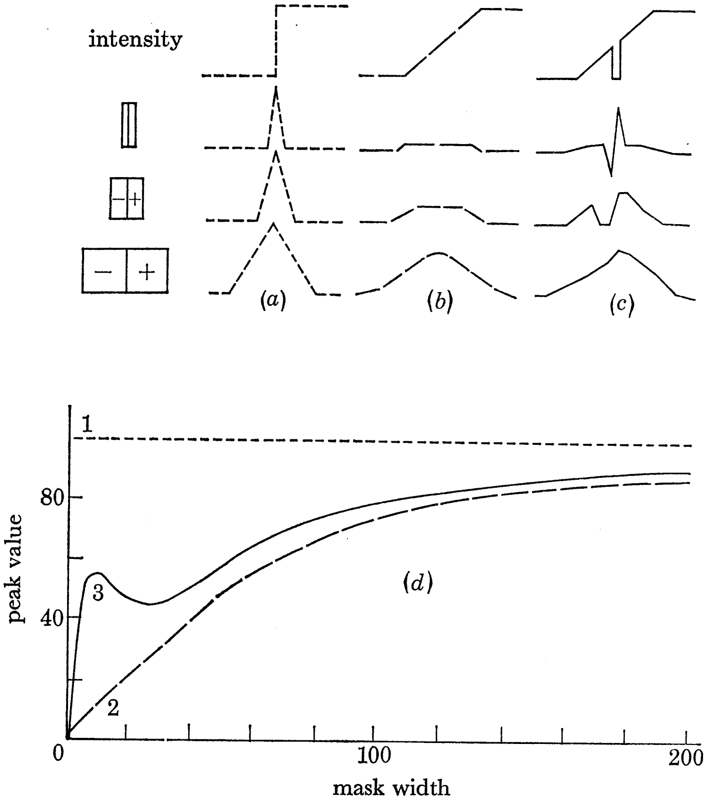


FIGURE 1. Selecting the appropriate mask size from which to compute the description of an intensity change. The figure illustrates the convolution of 'edge-shaped' masks of three sizes with different intensity distributions. The masks are shown to scale on the left, and the widths of their panels are 10, 25, and 60 units. The three intensity distributions are a step function (*a*), a function that increases linearly over 100 units and is constant elsewhere (*b*), and a step change 10 units wide superimposed on the linear one (*c*). Convolutions with each of these distributions are exhibited opposite each mask. In (*d*), the peak height that occurs in each convolution has been plotted against mask size for each intensity distribution; trace 1 corresponds to distribution (*a*), trace 2 to distribution (*b*), and trace 3 to distribution (*c*). The selection criterion chooses a mask size if it corresponds to a peak or to the left-hand end of a near plateau in the graph (*d*). Some distributions cause two mask sizes to be selected. Distribution (*c*) is one of these. The mask sizes selected for it are 10, and a value near 90. The curves in (*d*) are called *signatures*.

A competent edge-finder therefore needs to be sensitive to discontinuities in intensity and in intensity gradient, or (roughly) to measure the first and second derivatives of intensity everywhere. Approximations to these quantities may conveniently be obtained by convolving the image locally with 'edge-shaped' and 'bar-shaped' masks (see figure 2*a*). This follows from the fact that an edge-shaped mask measures an approximation to the local intensity gradient in a particular direction. A bar-mask may be thought of as composed of two adjacent edge-masks with opposite signs. It therefore measures approximately the local change in intensity gradient.

This argument defines the types of intensity change that are to be detected, but it is important to note that simply making the measurements is not enough. Almost every point in almost every natural image gives rise to a non-zero convolution value with almost every size and orientation of edge-mask. We therefore have to compute from this mass of data some symbol that represents a local piece of edge, and it is this symbol that will then stand in correspondence with a line segment in an artist's drawing. Fortunately, we can make a great simplification at this stage in the analysis. Provided that measurements are made with masks of two or more sizes, the positions and sizes of the peaks in the measurements provide enough information to compute the description of the underlying intensity changes. Furthermore, provided that a group of peaks is sufficiently isolated from other peaks, the other peaks may be ignored when analysing that group.

The reason for this is illustrated in figure 1, which shows the difference between edge-mask values obtained by using masks of three different sizes on a step change in intensity (figure 1*a*), and on a gradual change (figure 1*b*). The results are analogous to the power spectra of the different kinds of edge. Step changes are 'seen' equally well by all sizes of mask. Gradual changes are seen increasingly faintly by edge-shaped masks whose dimensions are smaller than the distance over which the intensity change is taking place. Figure 1*d* shows this effect in graphic form by plotting the maximum (absolute) edge-mask value against the mask width. Trace 1 arises from the step change (figure 1*a*), and trace 2 arises from the gradual intensity change (figure 1*b*). A good estimate of the spatial extent ('fuzziness') of an edge may be made by finding the mask size at which the edge-mask response starts to diminish. Accordingly the following criterion is used.

Selection criterion: mask size s is selected at point P in the image whenever (*a*) masks slightly smaller than s give an appreciably smaller peak at P, and (*b*) slightly larger masks give a peak that is not appreciably larger.

For some intensity distributions, more than one mask size will satisfy the selection criterion. For the distribution shown in figure 1*c*, the criterion is satisfied by $s = 10$ and $s = 80$ to 100 (depending on the algorithm that interprets 'appreciably'), as can be seen from trace 3 of figure 1*d*. Such a distribution would give rise to three assertions, a sharp negative edge close to a sharp positive one, and a fuzzy positive edge that encompasses the other two.

This shows one way in which the use of multiple mask sizes is important, but there is another reason which is nearly as important. It is that where a faint edge exists in the image, it is frequently impossible to tell from a single record which of the peaks are important, and which are due to noise. Matching peaks obtained by using different sizes of mask greatly aids the separation of signal from noise.

(The algorithm to which this leads is similar to the non-linear technique described by Rosenfeld & Thurston (1971).) The difference lies in the use to which the algorithm is put. Rosenfeld & Thurston used it for detecting texture boundaries at which the average grey-level change was small compared with the contrast occurring within each texture. To achieve successful results they required that measurements from masks of all sizes be available at all points in the image. (Note that unlike spatial frequency, the denser the measurements, the more information one has. If measurements are made at every point and sufficient information is available about the boundaries, a finite intensity array is completely recoverable from its convolution with any edge- or bar-shaped mask that is not too large.) In the present theory, texture boundaries are

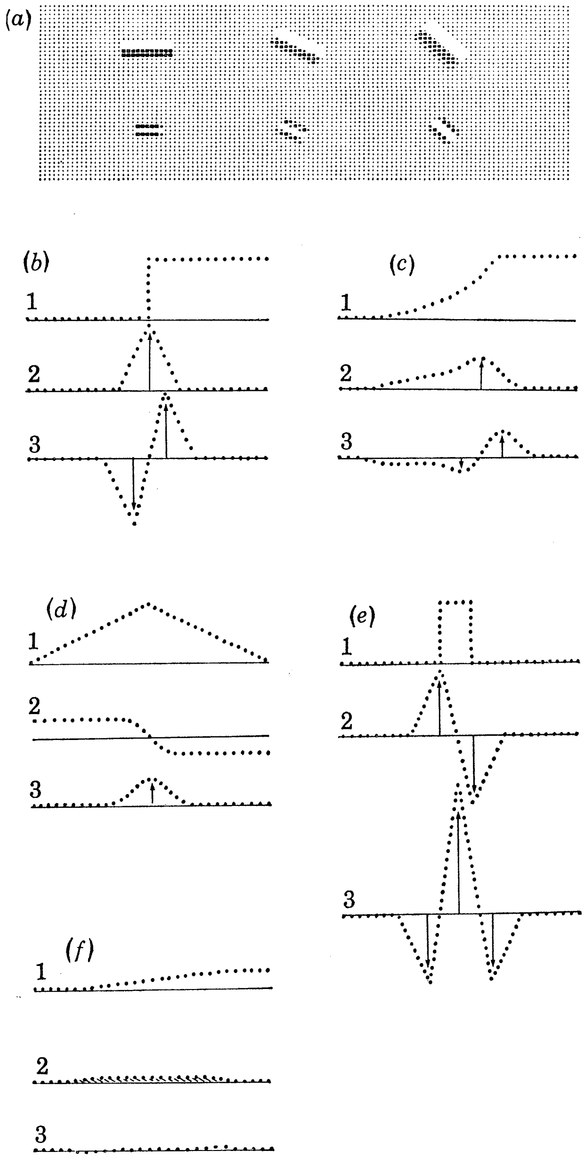


Figure 2. Classifying the grey-level intensity changes present in an image. Examples of the edge- and bar-masks that were used appear in (*a*). The text classifies the possible configurations of peak patterns in edge- and bar-mask convolution profiles, and this classification is illustrated by (*b*)–(*f*). Some typical intensity profiles are shown and marked with a 1; the corresponding edge-mask profiles (first derivative) are marked with a 2, and bar-mask profiles (second derivative) with a 3. The classes are EDGE (*b*), EXTENDED-EDGE (*c*), BAR (Mach band) (*d*), LINE (*e*) and SHADING-EDGE (*f*), which is the term we use for intensity changes whose spatial extent is large compared with the available mask sizes. Intermediate descriptions are used when the processor fails to find sufficient peaks to determine the edge type. No claim is made that this classification is the only possible one, although it is held that both first and second derivatives of intensity need to be taken into account at this stage.

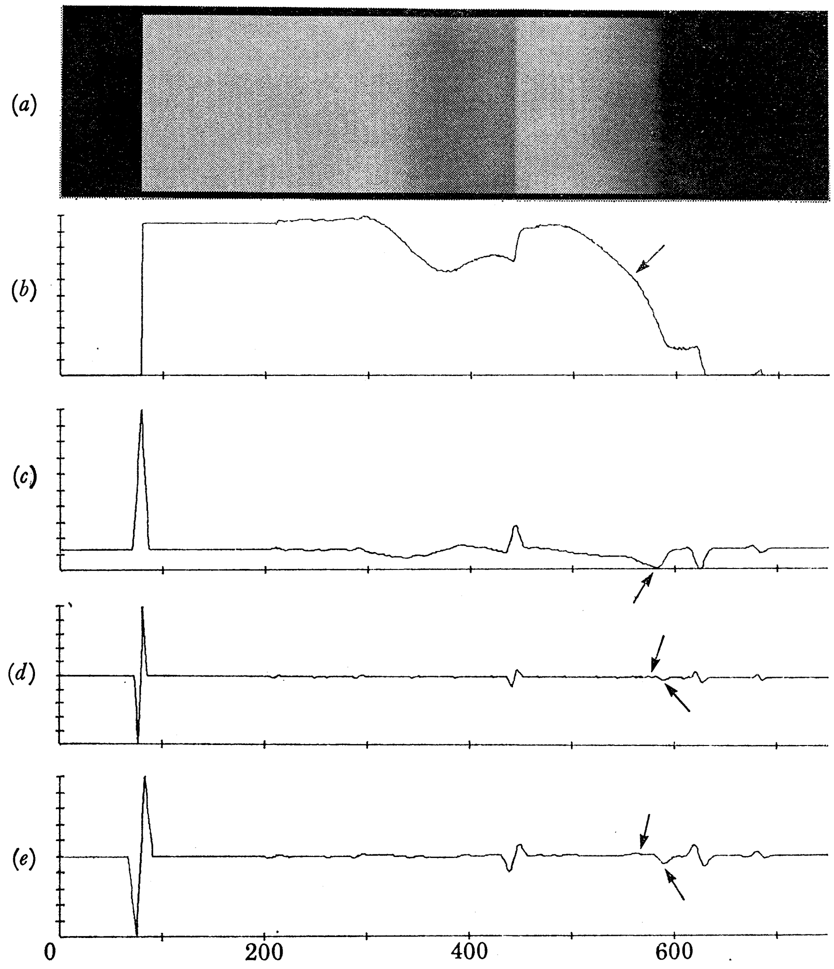


FIGURE 3. The intensity distribution exhibited in (*a*), whose profile appears in (*b*), was obtained by illuminating a curved piece of white paper from one end, and viewing it from above. Its description, computed by using an edge-mask of panel-width 8 (*c*), and bar masks-of panel-widths 4 (*d*) and 8 (*e*), is as follows:

EDGE (POSITION 80) (CONTRAST 136) (FUZZ SHARP)
EDGE (POSITION 212) (CONTRAST 3) (FUZZ 4)
EDGE (POSITION 292) (CONTRAST 2) (FUZZ SHARP)
EDGE (POSITION 435) (CONTRAST −3) (FUZZ 4)
EDGE (POSITION 444) (CONTRAST 25) (FUZZ 5)
EDGE (POSITION 464) (CONTRAST 2) (FUZZ 4)
EDGE (POSITION 490) (CONTRAST 1) (FUZZ 4)
EXTENDED-EDGE (POSITION 582) (CONTRAST −12) (FUZZ 9)
(the peaks giving rise to this edge are marked with arrows)
EDGE (POSITION 624) (CONTRAST −20) (FUZZ 6)
EDGE (POSITION 676) (CONTRAST 3) (FUZZ 4)
EDGE (POSITION 684) (CONTRST −4) (FUZZ 4)
SHADING-EDGE (POSITION 570) (CONTRAST −14) (WIDTH 67)
SHADING-EDGE (POSITION 391) (CONTRAST 4) (WIDTH 36)
SHADING-EDGE (POSITION 339) (CONTRAST −8) (WIDTH 73)

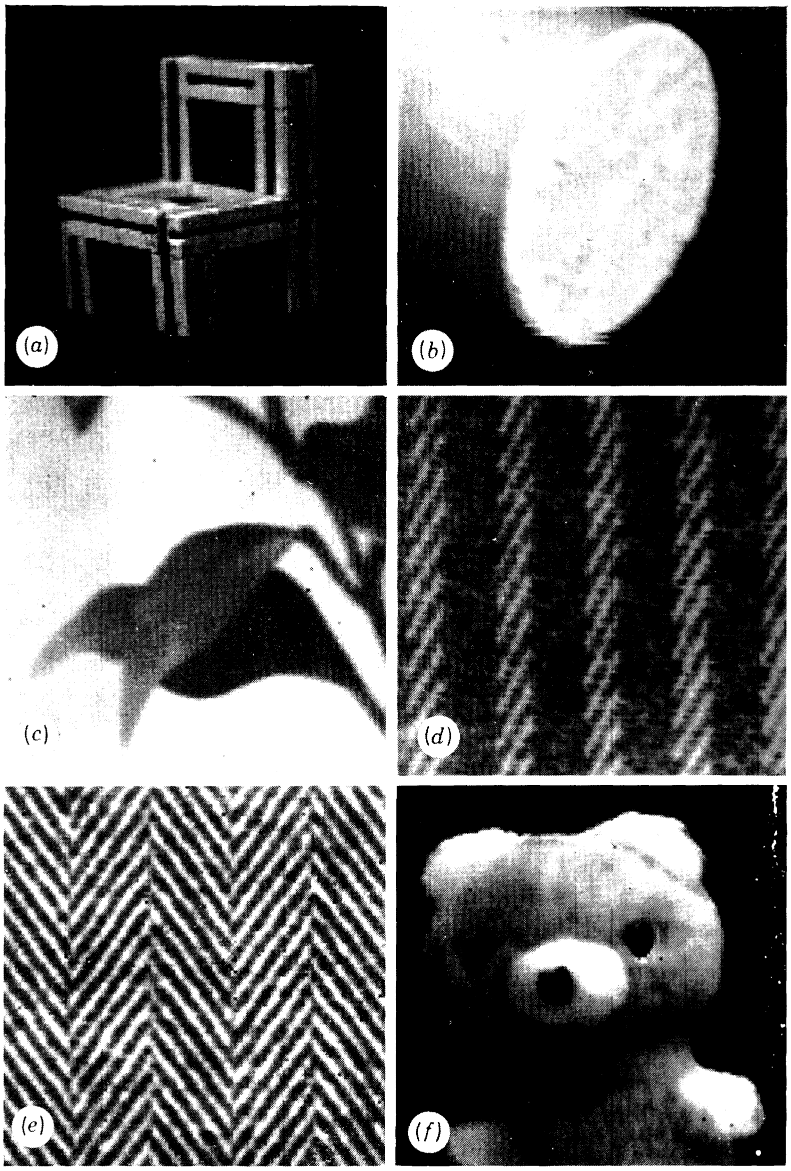


Figure 4. This figure provides a high quality reproduction of the six images discussed in the text. (*a*) and (*b*) were taken with a considerably modified Information International Incorporated Vidissector, and the rest were taken with a Telemation TMC-2100 vidicon camera attached to a Spatial Data Systems digitizer (Camera Eye 108). The full dynamic range from black to white is represented by 256 grey-levels. The images reproduced here were created by an Optronics P150ohPhotowriter from intensity arrays that measured 128 elements square. This size of intensity array corresponds to viewing a 1 in square at 5 ft with the human retina. The image of the period at the end of this sentence probably covers more than 40 retinal receptors. The reader should view the images from a distance of about 5 ft when assessing the performance of the programs. In the interests of clarity, these intensity arrays have been displayed in two other ways (where helpful). They have been printed on a Xerographic printer using a font of 16 grey-levels; and they have been displayed as a three dimensional graph, in which the *z* coordinate represents intensity. These displays appear in the figures.

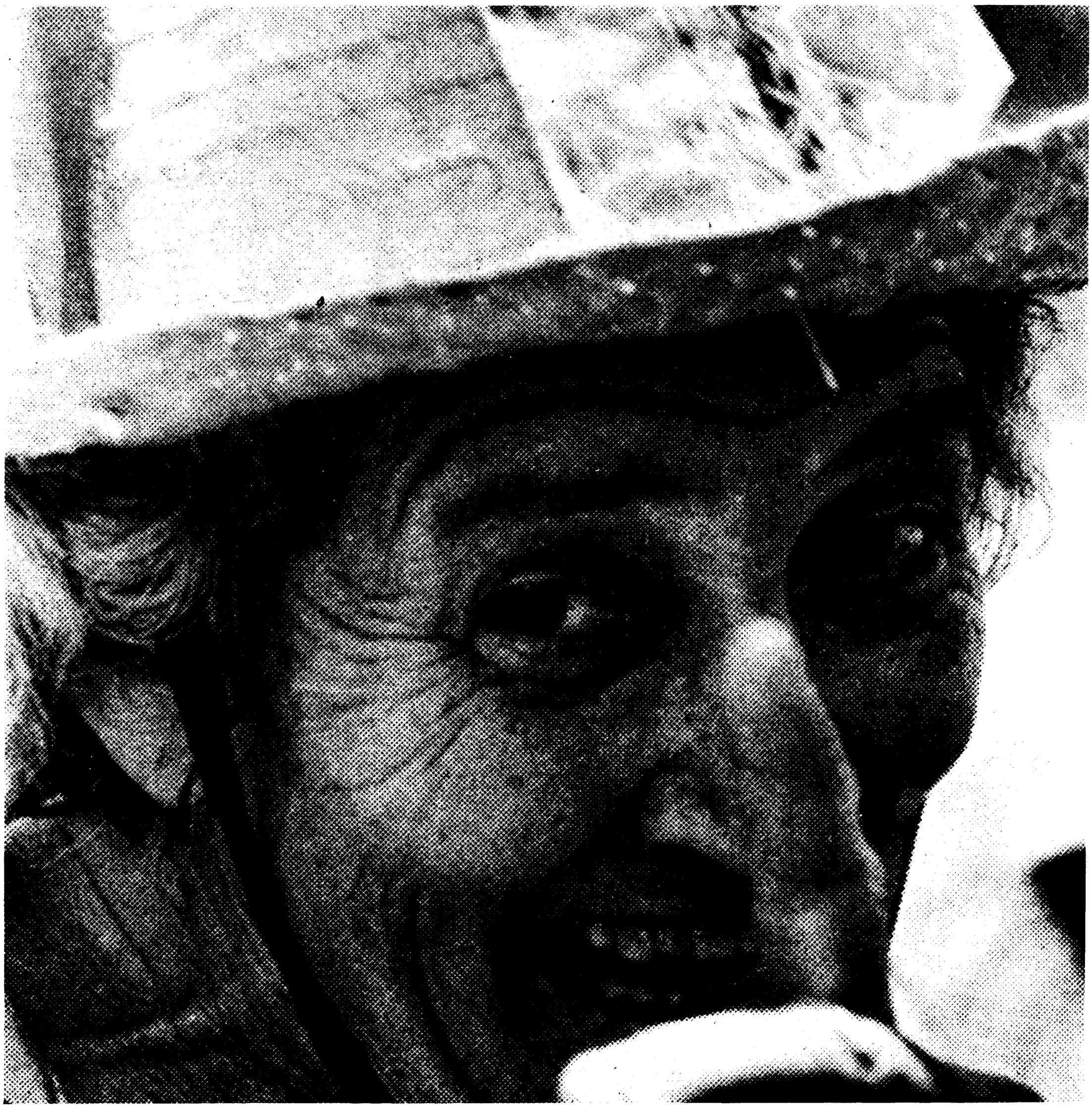


FIGURE 14. About two people in three fail to perceive the original of this image correctly the first time. The failure is caused by the accidental alignment of the subject's forefinger and nose. This failure shows that simple local processes are important during the analysis of an image, and that delivery by them of an incorrect grouping is not a normal event. This is good evidence against the hypothesis that early visual processing is designed around a failure-driven control structure. The fact that one does not make the same mistake a second time shows that some downward-flowing information can affect early processing. Only a small amount would be required to prevent recurrence of the error.

detected by other means, and the algorithm is used simply to obtain a measure of the spatial extent of an intensity change. Hence unlike Rosenfeld & Thurston's application, the distance between measurements can decrease as the size of the mask increases without weakening the technique.)

The process of computing the description consists then of four operations: (1) find and match peaks in the measurements obtained from the convolutions of the image with different sizes of mask; (2) select the relevant peaks using the selection criterion; (3) separate the peaks into isolated groups; and (4) parse the local configuration of peaks into a descriptive element. A small number of classes of peak configuration suffices to cover the cases that can actually occur, and they are illustrated in figure 2. The figure shows typical combinations of peak patterns that occur in the outputs from edge-mask (middle records of each triple, labelled with a 2) and from bar-mask (lower records labelled 3) convolutions with standard intensity profiles (upper records labelled 1). Examples of the masks that we use appear in figure 2*a*. The descriptor EDGE is used when two peaks of about equal and opposite signs occur together in the bar-mask record (figure 2*b*). If one bar-mask peak is considerably smaller than the other, the edge is classified as an EXTENDED-EDGE (figure 2*c*). Extended-edges are common where a convex boundary is illuminated from one side. Figure 2*d* shows an intensity gradient edge, and figure 2*e* corresponds to the presence of a thin LINE such as can occur in the highlight from an object's edge, or a very thin pencil stroke. Finally there are edges that begin and end gradually, and extend over a relatively large distance; these are classified as SHADING-EDGES (figure 2*f*). In addition to descriptors of edge type, one can measure an edge's CONTRAST, POSITION, ORIENTATION, and FUZZINESS. This last parameter characterizes the spatial extent of the edge.

Figure 3 gives an example of an intensity distribution that has been described by this process, and the legend explains which mask convolutions were used. One of the assertions has been traced back to the convolution profiles, and the arrows point to the peaks that gave rise to that particular assertion. The low-level vocabulary that is used in our present system is not intended to be definitive, but some claim is made to the effect that it is a good example of the genre, because it rests on the correct measurements, it has sufficient expressive power to describe most kinds of shading adequately, and the method is simple and works reasonably well.

Extension to two dimensions

The method may be extended to two dimensions by carrying out the analysis simultaneously at several different orientations. It is preferable to use orientation-dependent measures for making the initial measurements, for reasons that are illustrated by figure 5. The image (figure 5*a*) of a chair (128 points square), whose half-tone image is figure 4*a*, plate 1 and whose intensity distribution is shown in figure 5*b*, has been convolved with 'corner-shaped' masks. The results appear in figure 5*c* and *d*, but can the reader confidently distinguish the corners from these measurements? The reason for the failure is that the transform inverse to that produced by a corner-shaped mask depends critically on the boundary conditions that obtain. Any method that computes a corner *assertion* is saying something about this inverse and so must take enough information into account at each point to satisfy the dependence on boundary conditions. This extra information may be supplied by looking at the results of the corner mask at neighbouring points or by looking at the results of some other measurement taken in parallel; the important point is that the computation is not a trivial one and it has to take these extra factors into account.

The way to avoid the difficulty is to make the masks so orientation-dependent that they push the problem back into one dimension. To take account of the boundary conditions associated with edge- or bar-shaped masks, one needs to compare quantities in only two directions, rather than in all directions round a point. This makes it inherently simpler to compute the primal sketch from measures obtained with such masks, and it is why we use them. Notice that this argument is independent of presumed properties of the image. It is not impossible to compute

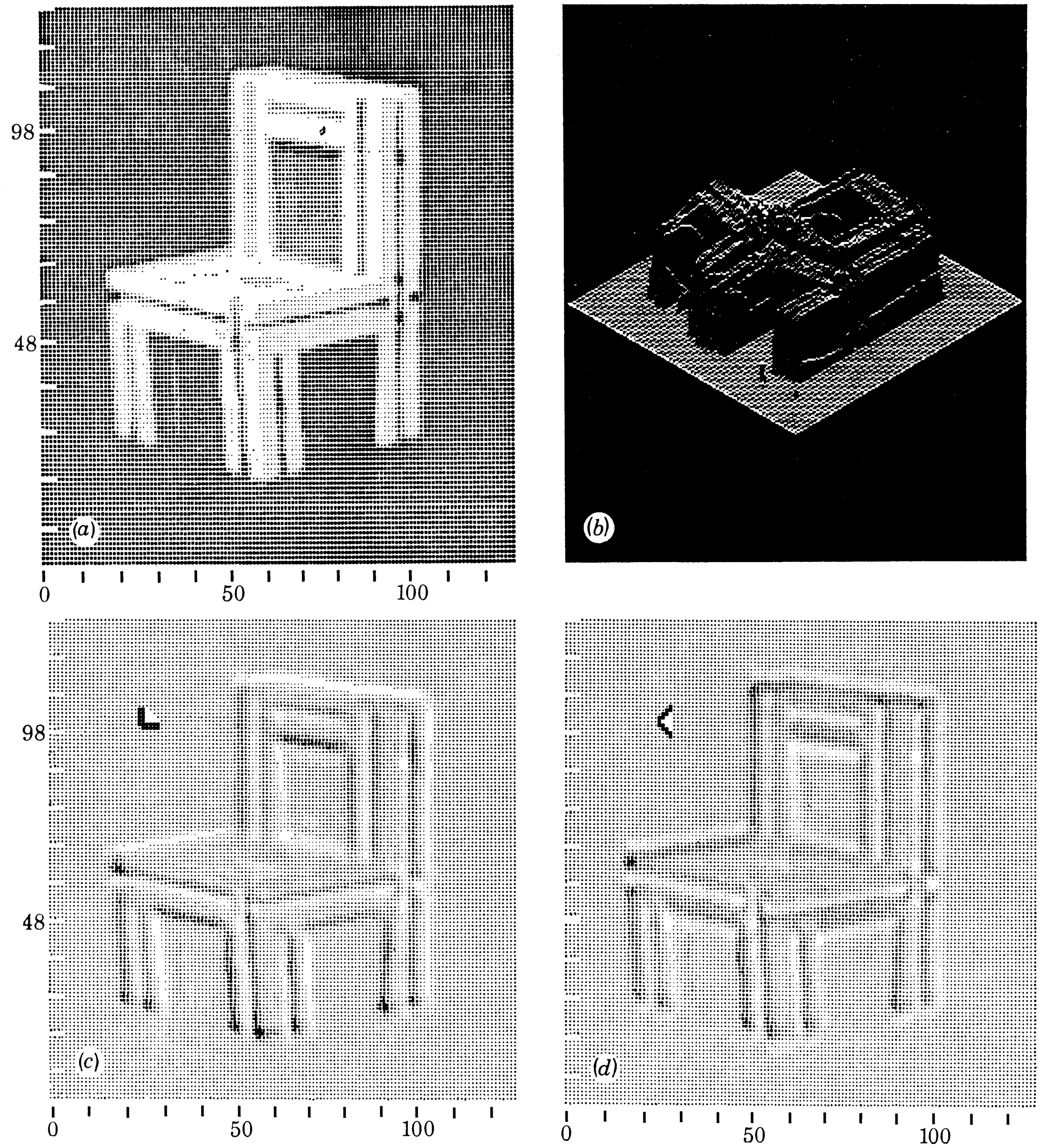


FIGURE 5. The image of the chair whose half-tone representation is given in figure 4*a*, has been printed in a 16 grey-level font in (*a*). A three dimensional intensity map (height = log intensity) appears in (*b*). This image has been convolved with two 'corner-masks' (*c*) and (*d*). Detecting corners from such measurements alone is not an easy task. This illustrates why it is difficult to compute a description of an image directly from measurements that are not directionally selective.

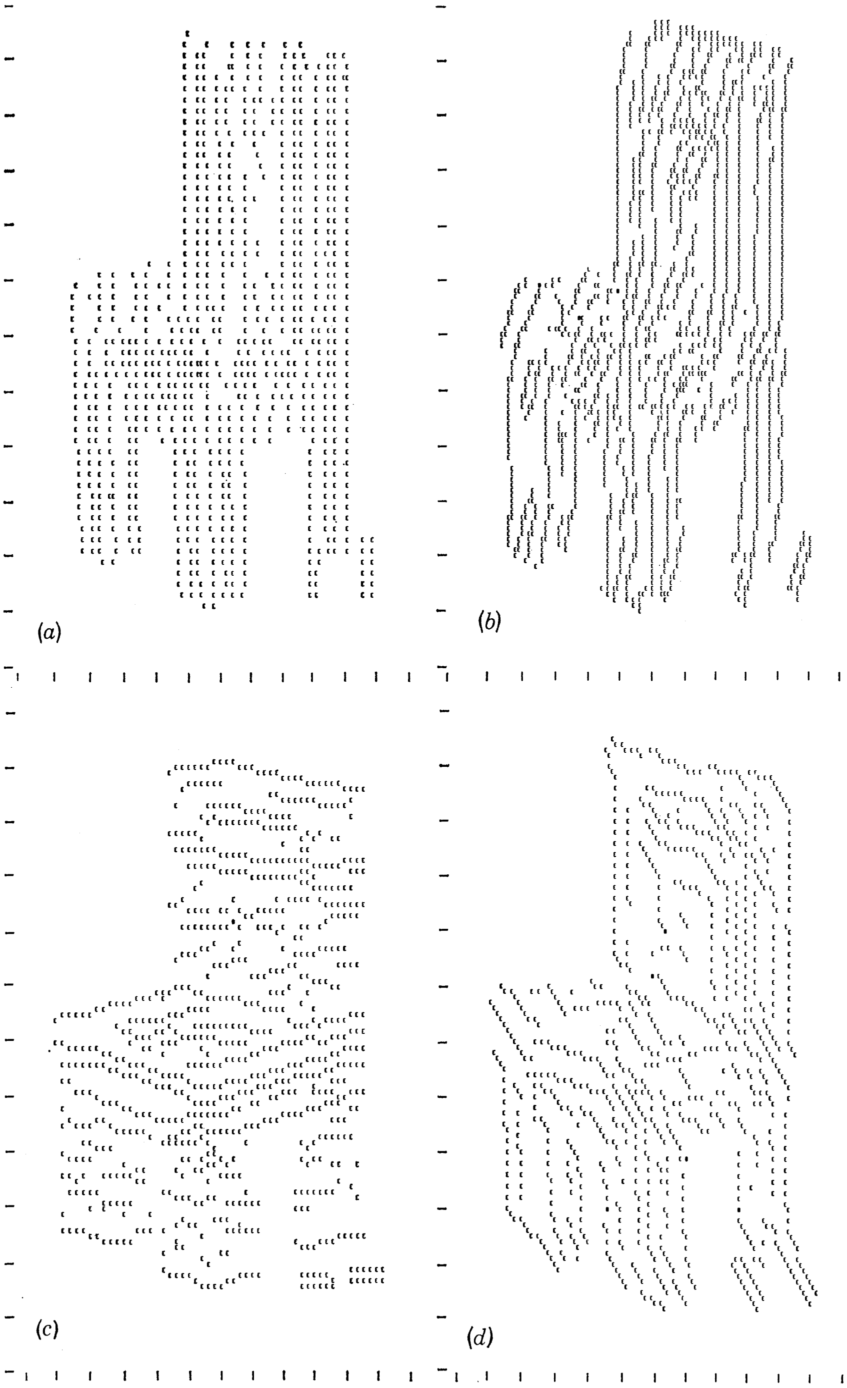


Figure 6. The first step in computing the primal sketch of the image chair is to compute a description of the grey-level changes at each of eight orientations. The results of doing this at four orientations are shown here. The orientations are arranged clockwise from the vertical (*a*), 22.5° to the vertical; (*b*) horizontal; (*c*), and 45.0° to the horizontal. The descriptions were obtained by scanning every other line perpendicular to the orientation of the masks. Each division on the axes represents ten image elements. Two sizes of bar-mask and one edge-mask were used. This included a bar-mask of panel-width 2 and length 10, in addition to the masks shown in figure 2*a*. Each of the letters in each figure represents an assertion like that given in the legend to figure 3. The axes are marked at multiples of 10 picture elements.

the primal sketch from measures that are not directionally selective, but a persuasive case would have to be made for choosing them.

Combining orientation-dependent descriptions

The number of different orientations at which the analysis needs to be carried out is fixed by the first stage at which the local assertions are glued together. The sensitivity of the masks is not so important, as we can see by calculating their orientation tuning curves. The ratio of panel-length to panel-width in the masks that we use is about 5:1 (figure 2*a*). If such a mask is rotated about a step-change edge, the angular distance between the maximum response and $1/\sqrt{2}$ of the maximum is about 35°; so their natural tuning curves are very broad.

Much more critical is the flexibility with which individual elements are combined to form assertions about small edge-segments. This process is the beginning of the grouping phenomena that seem to be central to early visual processing, and designing them has been the main stumbling block in writing competent edge-detectors. One of the best of them (Horn 1973) requires that lines should have length 20 before evidence of their existence is accepted as compelling. It was designed this way because if substantially shorter elements are accepted, a large amount of 'noise' appears in the output. Blobs and blotches, common in textured images, often give rise to elements that are shorter than this, so ways have to be found of dealing with the noise.

Figure 6 gives some examples of the data with which one has to deal. This shows the primary analysis of CHAIR at the vertical (figure 6*a*), 22.5° to the vertical (figure 6*b*), horizontal (figure 6*c*) and 45.0° to the horizontal (figure 6*d*). For each mask orientation, the image has been scanned along every other line perpendicular to the mask, and every point along each scan line was considered. We have to use a fine scan because the smallest masks used were so tiny. Each symbol E in figure 6 represents an assertion like that given in the legend to figure 3. With this scan, it is sufficient to use a primary grouping that operates independently along eight orientations 22.5° apart. The grouping requires that the types of adjacent primary assertions (represented by the E's) should roughly match (for example, EDGE matches EXTENDED-EDGE but not LINE), and that the relative positions of the two assertions should be appropriate. Edges whose orientations lie midway between two scanning directions are sometimes found by both neighbouring scans, which shows that eight orientations are sufficient at this stage. Some technical problems have to be dealt with before this process will work successfully, but they are too minor to be treated here (see Marr 1976*a*).

By the time the primitive elements have been assembled into straight edge-segments, evidence that they originated from eight scans has almost evaporated. It is advisable not to quantize the orientations of the glued edge-segments, because doing so can cause confusion between a straight line and one containing many small kinks. It is however possible to devise a discrete representation system for the segments, in which a segment of a given orientation is represented by linear interpolation between fixed, standard orientations. Most schemes of this sort require some mutual 'inhibition' between carriers of neighbouring components in order that the contrast of the intermediate edge should be represented linearly (see Marr 1976*a*). Such inhibition arises for purely representational reasons. The main force behind the initial gluing process is the consistency relations between nearby primitive elements.

Nevertheless there turns out to be a need for competition between scans at different orientations, that arises for reasons which are intrinsic to the analysis not just from a representational convenience. The surprising point is that the competition is required not between segments at nearly adjacent orientations, but between ones that are nearly perpendicular.

Figure 7 illustrates the problems that arise. The image of the end of a rod (figure 4*b*, figure 7*a*, *b*) was first operated on at eight orientations with the process described in the last section. Next, these local assertions have been glued along directions nearly parallel to the masks from which they were obtained. Each edge-segment in figures 7*c* and *d* represents several of the E's of the type shown in figure 6, and the database records all of the parameters associated with each segment. Quantities like the edge type, contrast and fuzziness are specified at intervals along the longer segments, since they can change along them. The longer segments should properly be regarded as a sequence of collinear short segments. In a full vision system, discontinuities of binocular disparity or motion along such an edge could still prevent the assembly of its sub-segments into a single unit.

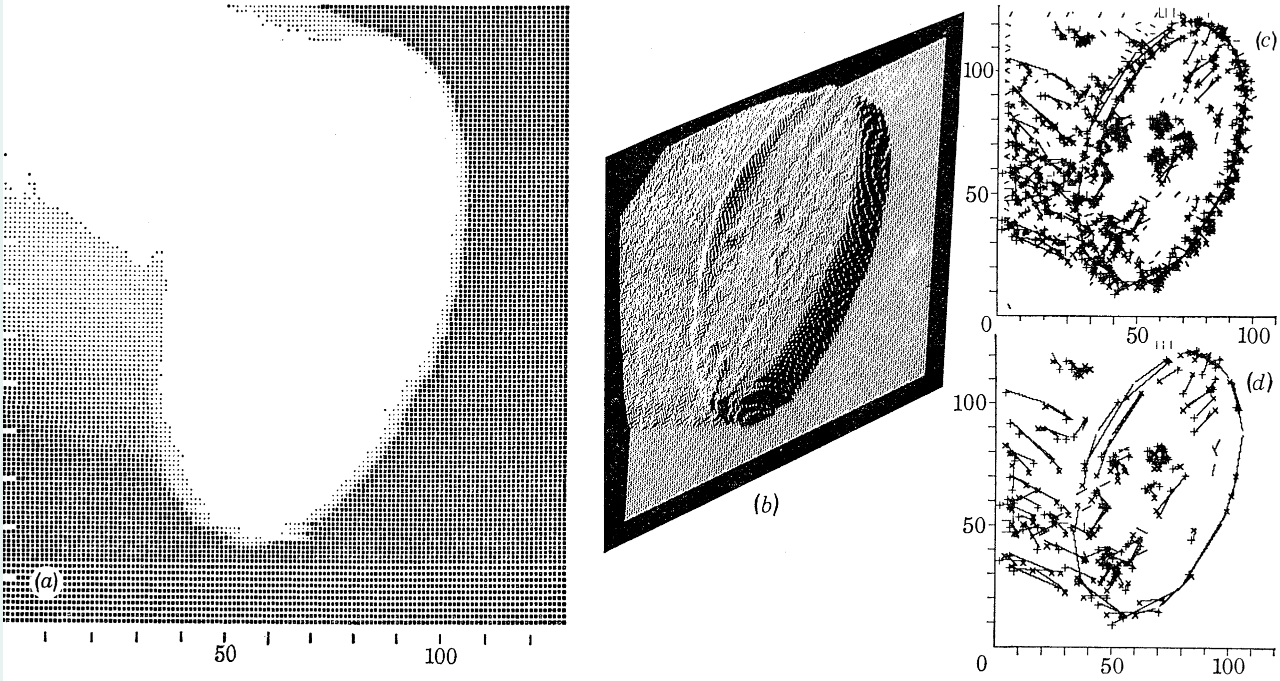


Figure 7. The second step in computing the primal sketch. After the intensity changes have been described independently at each of 8 orientations, and after local linear assembly of these descriptions has taken place, the eight descriptions are combined. This process is illustrated here for a particularly simple image, of the rod whose half-tone representation appears as figure 4*b*, plate 1. The printed version of this image appears as (*a*), and the intensity map as (*b*). The results are combined to give the data shown in (*c*). Each tiny line segment corresponds to two or more individual assertions (like those illustrated in figure 6), and a summary of the information associated with each of those assertions (as in figure 3) is made at intervals along each segment. Only information about the positions of the segments and about the precursors of termination assertions (shown as crosses) can conveniently be represented in a diagram; this can give a misleading impression of some items in the primal sketch. For example, many of the lines on the curved part of the rod on the left of the image arise from shading edges. They describe the gradual intensity changes that take place there, and should not be thought of in the same way as sharper edges. Short noise elimination then takes place to give (*d*), which gives a fair idea of the messiness of the uninterpreted primal sketch.

The feature of the data that is relevant to inter-orientation competition is the abundance of short segments roughly perpendicular to the primary edge (figure 7*c*). These are caused by a combination of local noise, the image tesselation, and irregularities in the image. They occur in every image that we have processed. In dealing with them, one cannot dismiss in a cavalier manner all very short segments: tiny 'blobs' in the image also give rise to them, as can be seen from the same image at coordinate (73, 75). But a 'small' element like this can be ignored if (*a*) it crosses a 'long' element, and (*b*) its contrast is less than that of the item it crosses. Figure 7*d* shows the results of removing small noise elements using this criterion. Occasionally, two small noisy segments can accidentally become aligned, creating a longer noisy segment. These are eliminated in the same way.

The crosses in the figure (sometimes rotated to avoid alignment with the edge segment to which they are attached) signify that the contrast of a directed segment changes rapidly at that point, possibly becoming zero. They are the precursors of assertions about the presence of terminations, and may be thought of as identifying the exact position of a termination if one exists near there. The problems that arise in obtaining them are dealt with elsewhere (Marr 1976*a*).

One other item of note in computing the primal sketch is the question of detecting local, small blobs. Figure 7*d* at coordinate (73, 75) shows how they appear, and in fact we make small blobs a primitive element of the primal sketch, together with their associated contrast, and the sizes and orientations of their major and minor axes. The defining criterion for when an image item is small enough to be called a blob or a line is that it should be indivisible. This occurs when one of its dimensions is comparable with the resolution of the analysis of the image at that point (about 5 image elements in length). Finding blobs from the glued assertions depends a small amount on elegant programming, and a large amount on brute force (Marr 1976*a*).

Some consequences

As a model of the information-processing that is performed in area 17 of the monkey, these ideas have one main consequence whose disproof would destroy the theory. It is that the direct output of a linear simple cell is not available centrally. Its signal is used to create an assertion about the presence of an edge, and that assertion is what is available. Creating the assertion is an act of computation – a simple one, since it involves little more than peak matching, applying the selection criterion, and the classification of a peak configuration; but it is an act of computation nonetheless. The main point is that this has to go on, and one should therefore be able to find experimental evidence of it.

A consequence of this view is illustrated in figure 8. Suppose that an image contains two small close blobs. These blobs give rise to measurements by a number of sizes of mask – some small ones represented by the tiny line segments, and some large ones, like the one that is illustrated. One's *a priori* inclination might be that a large 'line-detector' would fire, and that this would have something to do with seeing the two blobs. This view amounts to supposing that simple cells write directly into a feature-point array that is freely available to subsequent processes. But if our theory is correct, although the large 'simple cell' may indeed fire, its measurement will not be used to compute the description of the two blobs because their sharp boundaries cause the associated intensity change to be described from peaks in the small masks (by the selection criterion). The selection criterion (figure 1*d*) will cause the description to be computed from the smaller masks unless the blobs are severely defocused.

Another interesting point is that we fail to 'see' Abraham Lincoln in L. D. Harmon's coarsely sampled and quantized image of him (reproduced by Julesz 1971, p. 311). If measurements from linear simple cells were freely available to later processes, and if we were able to select them by receptive-field size, we would presumably be able to interpret that image without physically defocusing it. According to the present theory, the mask size used to compute the description is chosen by the selection criterion. This is consistent with Harmon & Julesz's (1973) finding that noise bands spectrally adjacent to a picture's spectrum are most effective at suppressing recognition, since these have most effect on mask response amplitudes near the important mask sizes. Furthermore, because two peaks in the graph 1 *c* would cause the algorithm to create two local edge assertions (with different degrees of fuzziness), it also explains why removal of only the middle spatial frequencies from such an image leaves a recognizable image of Lincoln behind a visible graticule (figure 1 *d* of Harmon & Julesz 1973).

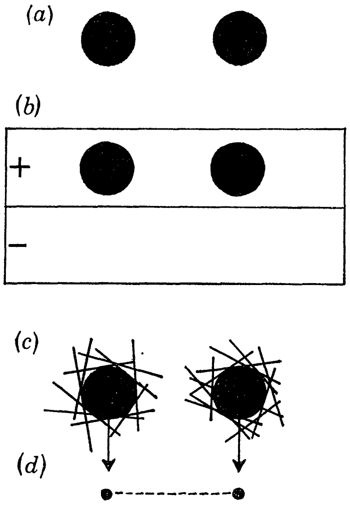


FIGURE 8. The difference between the primal sketch and a feature-point array is brought out by an intensity distribution like (*a*). A measurement taken with a large mask (*b*) could generate a feature-point, but it would not be used in the computation of the primal sketch. This is because the sharp contrast changes operate through the selection criterion to force the description to be computed from small masks like those shown in (*c*). The final description is of two blobs, which define 'place-tokens' (*d*).

The structure of the raw primal sketch as it is first delivered from the image may be summarized as follows:

PS1. The primary visual processor delivers a symbolic description of the intensity changes present in an image. This description uses the following primitives to describe intensity changes:

(i) various types of EDGE,
(ii) LINES, or thin BARS,
(iii) BLOBS.

The items (i) and (ii) have been assembled into straight segments, and short noise elimination has occurred.

PS2. The following items are computed and bound to each element of the description.

(i) ORIENTATION, of an edge, line or bar; of the major axis of a blob or group.
(ii) SIZE – length and width if both are defined, diameter if major and minor axes are equal or undefined,
(iii) local CONTRAST,
(iv) POSITION,
(v) TERMINATION POINTS.

What drawings tell us

The second step of the argument depends on our ability to interpret simple pencil drawings that lack semantic content. By examining suitable examples, we can infer with some confidence that certain symbolic grouping operations must exist in our visual systems. In order to establish the principle that grouping processes sometimes exist, let us first take an extreme case. When one looks at figure 9*a*, there can be little doubt that some process is creating a circular contour joining the inner ends of the radial lines. The path of this contour is marked by an apparent change in brightness, less than but comparable to that observed in Kanizsa's (1955) illusory white triangle whose apices lie in the white sectors of three suitably placed black disks.

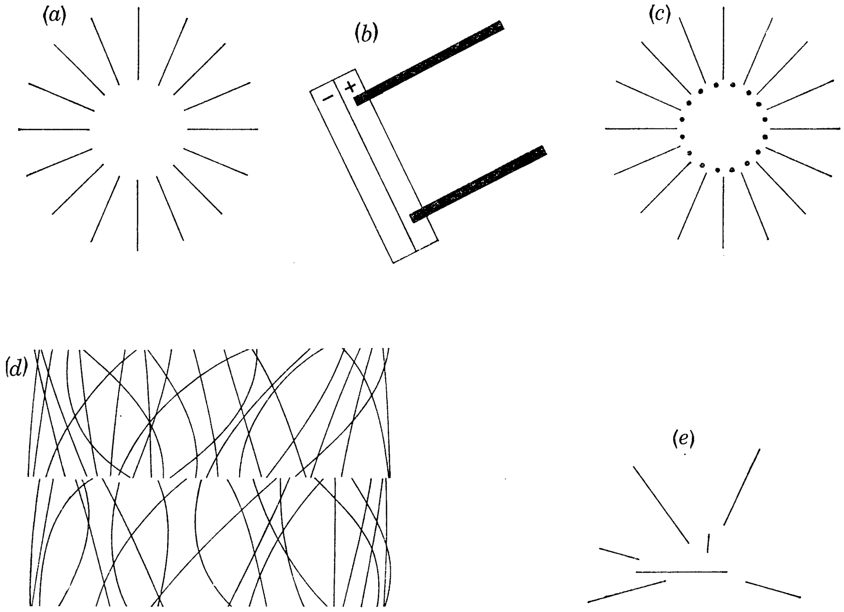


Figure 9. The illusory contour in (*a*) is somewhat similar to the Kanizsa triangle. It cannot be due to a simple cell in configuration (*b*), because the contour is still visible in (*c*). It cannot be due to a gestalt of the sun induced by radial lines, because the lines in (*e*) are radial, yet this is not readily apparent. A similar illusion is present in (*d*), suggesting that the apparent brightness of the inner disk reflects a default assumption about foreground–background contrast, rather than any high-level influence. The theory attributes the contour to local processes that join nearby ends of lines. Such processes are mechanisms of construction rather than mechanisms of detection.

In deciding how this comes about, we may distinguish three rival theories. (1) A local process operates to join neighbouring ends of lines. (2) The inner contour is constructed by some mechanism that relies upon the placing of an edge-shaped mask in the position shown in figure 9*b*. (3) The radial lines cause a 'Gestalt' of a 'sun' to be used for describing the situation. This very high-level concept then imposes the contour on the figure.

If (2) were correct, it would disprove the primal sketch theory, since it requires that a mask output value be identified in a simplistic way with an assertion about a contour. Figure 9*c* disproves (2) however, because the contour remains visible despite the presence of an intensity distribution that would remove or negate the mask values on which (2) depends. If (3) were correct, it would imply that downward-flowing information has a great influence on early processing – a view which runs counter to the second main thrust of the present theory. Theory

(3) assumes a sensitivity to radial lines. The lines in figure 9*e* are however also radial, and this is not immediately obvious.

The possibility remains that some combination of (1) and (3) is what really governs our perception of the figure. The important point is that the initial acquisition of the 'sun' concept probably relies on the mechanisms in (1). Once accessed, this Gestalt may influence the computations to the extent of deciding that the sun part is the foreground and is therefore slightly brighter, but such an influence determines only one bit of the final description. Figure 9*d* makes it unlikely that the particular 'sun' Gestalt has even this effect, since it provides a similar example in which 'ends-of-things' form a perceptually 'brighter' obscuring region. It is more likely that the relative brightness reflects a (context-sensitive) assumption about the sign of foreground–background contrast.

These examples establish that abstractly defined places in an image can be assembled into contours that have a definite perceptual existence, and that this operation probably precedes the access and application of higher-level concepts to the image. From a computational point of view, it is natural to think of the phenomena as occuring in two steps. Firstly, certain things in drawings can cause 'place-tokens' to be defined in some abstract sense. Secondly, place-tokens so defined can be grouped in various ways.

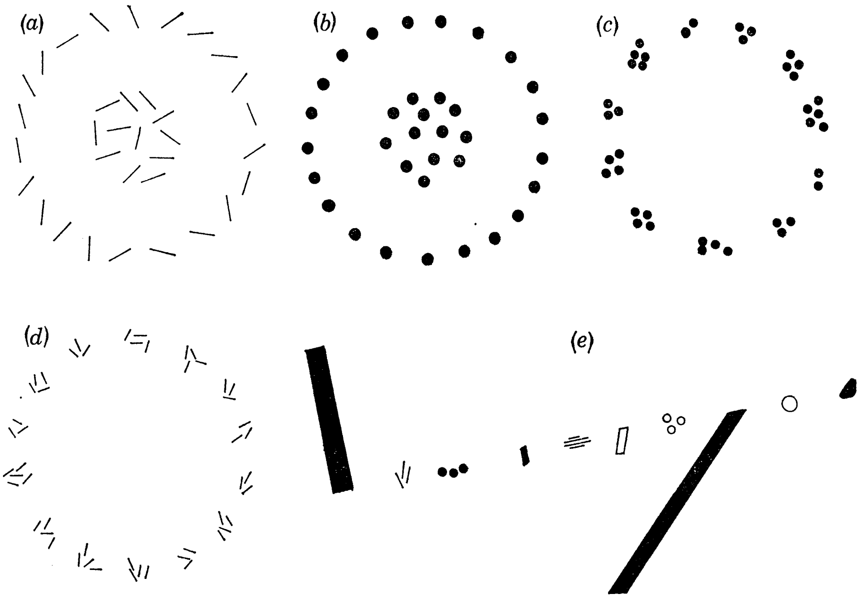


FIGURE 10. Place-tokens may be defined in an image in several ways, and may then be aggregated by certain standard techniques. Small lines (*a*) or blobs (*b*) may define a place-token. So may small collections of places (*c* and *d*). The definition and the grouping of place-tokens may be regarded as independent processes, because grouping does not depend on the way the place-tokens were defined. This is shown by (*e*), in which every sub-group is defined differently, yet the collinearity of all of them is immediately apparent. Information such as orientation may be bound to a place-token because it was intrinsic to the element that gave rise to it. Such information may be used to help grouping.

In how many ways may place-tokens be defined, and in what ways may they be grouped? We see from figure 10*a* that a short line may define a place-token, and from figure 10*b* that a small blob may also do so. The end of a line that is not too short, or of a blob with long major axis and short minor axis may also define a place-token. (The imprecision of the boundary between 'too long' and 'too short' is inconsequential, because near it, both definitions usually lead to the same groups. The boundary needs to be in the region of 0.5 to 1 degrees of arc at human foveal resolution.) Small collections of blobs (figure 10*c*) or of lines (figure 10*d*) may also be treated as a unit. Because of the variety of ways in which this may be done (figure 10*e*) it is probably implemented by the rule that a group of place-tokens may also define a place-token, rather than by different rules for groups of blobs, groups of lines, groups half of blobs and half of lines and so forth. Hence although place-tokens can be described and to some extent selected by properties of items at that place in the image, the grouping processes themselves read place-tokens and are insensitive to the particular way a place-token was obtained. The notion of a place-token is a good example of the principle of explicit naming, and the separation of the way in which a place-token is defined from the way in which it is grouped illustrates the principle of modular design.

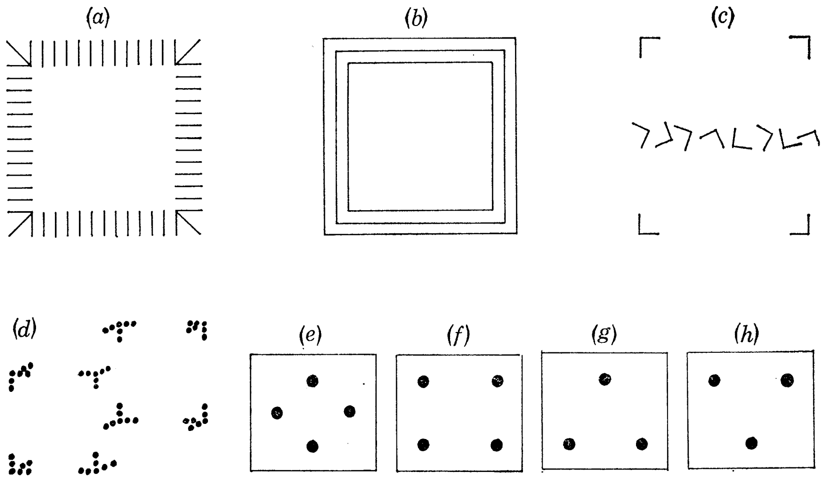


FIGURE 11. (*a*) and (*b*) give examples of groupings in which orientation is important. In (*c*), orientation is important for constructing the square, but not for perceiving the collinearity of the rotated 'L's' across the middle. Information about similarity of orientation is used if it can be, but it is not disastrous if it cannot be. (*d*) shows how the orientation of a small aggregate can be used to form a larger aggregate. Evidence like this suggests that the results of these primary aggregation processes are written into the same storage as the primal sketch. (*e*)–(*h*) give some examples of 'standard configurations' that we have found it useful to recognize. The reader will probably perceive them relative to a vertical axis. The VEE shown in (*h*) is used in figure 21*d*.

The recursive character of the definition of a place-token leads one to expect that the grouping processes responsible for them read and write into the same storage. Otherwise, one would have to maintain many copies of the storage and grouping processes, instead of just one. If only one copy is kept, two organizational rules must be observed. First, a priority system is needed that operates among competing processes such that (for example) very local groups usually take precedence. Interaction between rival local groupings is often necessary to arrive at a grouping satisfactory to them all. (In figure 10*e*, the local groups are formed before their organization into a line.) Secondly, when a grouping for a set of place-tokens is finally decided upon, only the token for the new group is subsequently visible to the grouping processes. The constituents of the new group are thereafter accessible only through the token for that group.

Grouping processes are sensitive to orientation, intensity (lightness), fuzziness, and various measures of the size of an item in the image, as well as to spatial proximity and collinearity. For example, orientation information may or may not be present (figures 10*a*, 10*b*, 11*a*, 11*b*), and if present, it may (11*a*, 11*b*) or may not (10*a*) be used. Indeed these two situations can occur in the same figure (11*c*). Combinations of spatial proximity and of similar orientation are often important.

We see from these examples that place-tokens can be grouped into regions directly, or into curvilinear assemblies that define regions by acting as their boundaries. The Gestalt psychologists were aware of these grouping phenomena (Wertheimer 1923). In addition to the region-defining facilities just mentioned, if the number of places involved is very small (less than 5 say), the places may form a standard, named configuration (see figure 11*e–h*) which is evidently described relative to an axis that is imposed on the figure, and whose default value is the vertical.

Separating figure and ground

Before the digression of the last section, we had reached the point of defining the raw primal sketch, and of showing how to compute most of the quantities in it. We also examined the primal sketch of a very straightforward image, of a rod. The primal sketch is rarely as simple as that, however. Figures 13, 18, 19 and 21 contain examples of the primal sketches of more complex images, and as one might expect, they are in general large and unwieldly collections of data. Furthermore, it is difficult to see how the complexity of the primal sketch could be an artifact of our particular choice of primitives: images really are complex in this way.

The unwieldy nature of the primal sketch creates what appears to be the main task of the next stage of visual information processing: how do we select regions that should be treated as unit forms by subsequent descriptive processes; and can this be done without complex interactions between the primal sketch and hypotheses about the nature of the forms that are being extracted? In perceptual terms, the computational problem that we must now address corresponds to distinguishing between figure and ground, and it is strongly related to the problem of texture vision (Julez 1971, see, for example, pp. 105 ff.). In neurophysiological terms, if area 17 roughly speaking computes the primal sketch, we come now to the problem that the next stage must solve.

We have now reached the core of the first part of the theory. We saw in the last section that certain computational facilities exist and are deployed during our reading of certain kinds of drawings. It is of course possible that their existence is no more than a happy accident, which fortuitously allows us to interpret the idle scribblings of the artistically gifted. The present theory was however founded on the observation that drawings and images appear surprisingly similar. It takes the view that the processes exhibited by the drawings of figures 9, 10 and 11 are not empty examples: the ability to perceive the envelope of a tree, a row of bushes, or even the border of a grass lawn can depend on such processes, and they are part of the reason why computer vision has had such problems finding object boundaries in the past. *A central assertion of this theory is that these grouping processes are available precisely because they are needed to help interpret the primal sketch; and furthermore that these symbolic processes, together with first order discriminations, operating recursively on the description in the primal sketch, are sufficient to account for most of the range of 'non-attentive' vision of which we are capable, within the class of images to which this article is restricted.* In other words, the extraction of forms and associated 'texture' discriminations are actually implemented by first order discriminations, together with a small number of grouping operations, acting repeatedly on the primal sketch of the image. We now study in more detail the grouping operations on which the second part of the theory depends.

Grouping techniques

The purpose of the grouping techniques outlined here is therefore to partition the primal sketch into unit forms, in a way that is useful for subsequent recognition. The important question concerns the extent to which hypotheses about the nature of a form need to interact with the processes that extract it. The issue is one of degree, not principle, since we shall show that some downward-flowing information may be necessary to complete segmentation. The demands of speed and fluency make it desirable to minimize these downward influences, and our main conclusion is that for most images, such influences affect only a small number of the decisions taken during grouping.

The most important guideline for the design of grouping techniques is the principle of least commitment. According to this principle, each step is irreversible. Hence only groupings that are reasonably certain may be made. This forces one to decompose the overall process into several steps, and to take advantage of as many cues as possible to help in the decisions that are made at each step.

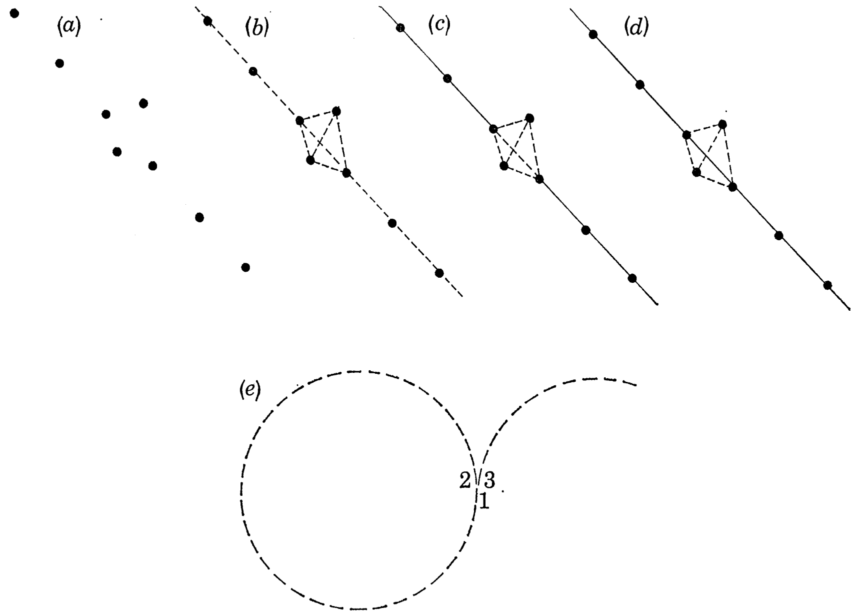


Figure 12. Examples in which semi-local and global constraints can influence local measures of preference during aggregation. (*a*) shows a set of place-tokens, and (*b*) illustrates the possible pairwise groupings that local neighbourhood analysis permits. The situation after the first pass is shown in (*c*). Information that was obtained from this pass about orientation combines with the equal spacing of the collinear tokens to make (*d*) the preferred linking on the second pass. In (*e*), the links between 1 and 2 and between 1 and 3 are evaluated as equally desirable on purely local grounds. The overall closure property creates a preference for the link that uses 2. In the primal sketch, the affinity between two elements is evaluated simultaneously along several dimensions. Considerations such as these can often cause a particular grouping to emerge as clearly preferable to any others.

Curvilinear aggregation

We define curvilinear aggregation to mean the assembly of place-tokens that contain an orientation into a group that preserves it. This type of aggregation is one dimensional rather than two, and the discovery and use of the appropriate local orientation is central to it. We shall see that one dimensional grouping processes are by far the most important kind. Two dimensional grouping seems to be necessary only locally; larger regions that are characterized by a texture predicate are best isolated by finding their boundaries.

Information that determines whether two items should be grouped comes initially from their primal sketch parameters and spatial dispositions. The primal sketch parameters are orientation, contrast, type (EDGE, LINE etc.), and fuzziness. Spatial information includes the distance between the nearest parts of the two items, and the relation between the orientations associated with the items and the orientation of the line joining their nearest parts.

Because of the principle of least commitment, the first stage of grouping combines two elements only if they match in almost all respects, are very close to one another, and if there are no other candidates. This typically reduces the number of groupable elements to about a third of the number present in the raw primal sketch. The second stage can then make use of extra information given by the first. Sometimes, the only extra clues are that some segments are now quite long (more than 20 image elements). Such segments almost certainly have some physical importance, and hence in the second stage it is safe to combine two such elements even if they fail to match on some parameters, provided that there are no other reasonable candidates in the vicinity. In some situations, the first stage will actually have introduced new information which can then be used by the second stage. For example, figure 12*a* shows a set of places that are to be aggregated, and the possible links between nearby places are shown dotted in figure 12*b*. By the second stage, an orientation parameter is present, and this, together with the equal spacing of the collinear tokens, makes the grouping shown in figure 12*d* the preferred one.

Some results of these two grouping processes are illustrated by the analysis of the image PLANT, which is exhibited because it raises several points of interest. Figure 13*a* gives the printed image whose half-tone representation appears in figure 4*c*, plate 1 and figure 13*b* shows its primal sketch. Figures 13*c* and 13*d* show typical segments obtained by the above processes. Notice the ragged nature of 13*d*; this is a common feature of the high resolution analysis of indistinct object boundaries. The local orientation of the raw primal sketch elements is preserved only roughly here.

Having exhausted all those situations in which aggregation takes place more or less by default, we turn now to the other technique that characterizes an application of the principle of least commitment, namely the rejection of relatively unlikely possibilities. The method is to set up a node for each of the ends of the segments that were delivered by the preceding processes, and to associate with each node a list of the nodes that could possibly match this one. Notice how this presupposes that each segment-end can be assigned an internal name (principle of explicit naming). Each of the possible matches is then evaluated independently along several dimensions, and possibilities that are graded relatively poorly by several methods of evaluation, and well by none, are struck out.

Our present implementation assesses the possible choices using measures of relative contrast, orientation, alignment or misalignment, distance, edge type, fuzziness, whether an item acts as a good intermediary between two segments that match very well, and whether a closed form would be created by choosing a particular segment. The idea behind this is straightforward. It has long been known to the Gestalt psychologists that in a line-drawing, each of these criteria can cause elements to be grouped together in a 'preferred' way (Wertheimer 1923). In the much richer environment of the primal sketch, there is frequently enough information available to apply to all of these criteria simultaneously. If most or all of them agree in selecting a particular grouping, one can be certain enough of its correctness to select that grouping irrevocably.

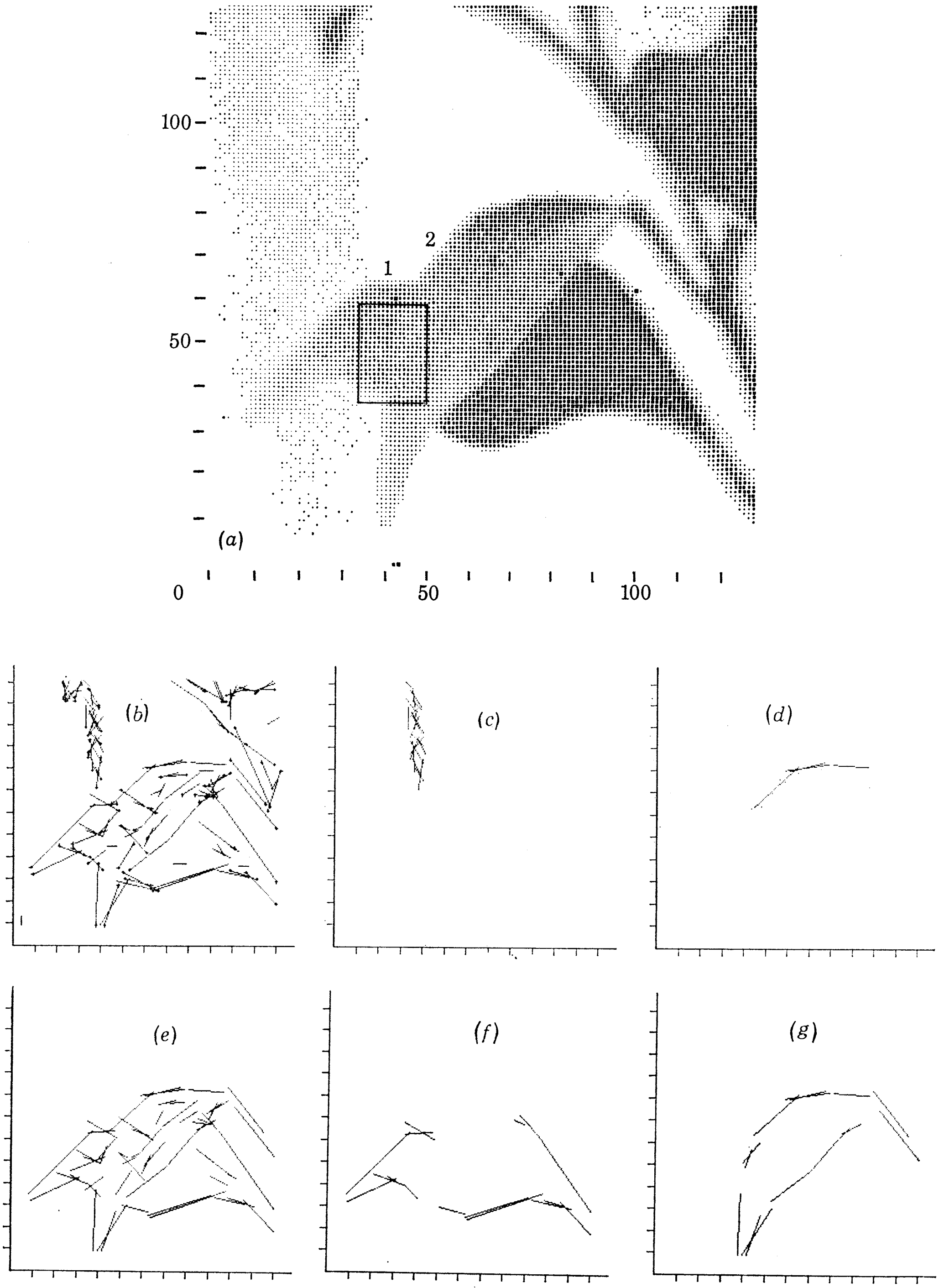


FIGURE 13. The image PLANT, whose half-tone representation appears in figure 4 *c*, has been printed in (*a*). The actual intensity values that occur within the superimposed rectangle have been set out in table 1. The spatial information from the primal sketch of this image is given in (*b*). Typical segments that arise from the first two stages in curvilinear aggregation appear in (*c*) and (*d*). The primal sketch does not contain quite enough information to separate the two leaves, and the aggregation techniques deliver the form (*e*). They have however almost succeeded in the separation. If one piece of information is added (that segment 1 does not match segment 2), the aggregation routines can separate (*e*) into (*f*) and (*g*).

TABLE 1

(The top table shows the intensity values for a small section of the image PLANT (see figure 12); the lower table gives the values of edge-mask convolutions over the same region. Only residual decay from the edge above this region is measurable. No general-purpose edge-finder could discern the edge of the nearer leaf in this part of the image.)

X = *Y*	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
58	171	169	167	167	166	165	166	164	167	171	171	174	174	175	173	171
57	168	168	168	167	166	167	167	165	169	168	174	176	175	175	175	172
56	168	167	167	165	166	166	167	167	168	170	178	177	176	174	174	173
55	168	168	165	169	167	168	167	165	168	175	177	177	175	175	172	171
54	169	170	167	169	169	168	163	166	172	169	174	173	175	178	173	173
53	171	169	170	168	169	168	169	168	168	170	175	173	175	177	178	176
52	172	171	170	168	169	169	167	168	173	172	173	177	174	175	178	176
51	172	174	171	170	166	168	167	168	172	172	172	177	179	172	175	175
50	171	167	176	169	170	169	168	169	171	172	174	174	173	173	174	178
49	174	172	173	173	173	174	171	171	172	174	172	172	172	169	173	173
48	173	173	173	176	178	172	171	174	174	173	175	175	175	173	173	171
47	173	175	178	173	173	171	171	175	175	177	178	175	174	173	175	178
46	178	175	174	169	173	175	177	175	177	177	174	175	176	177	177	174
45	173	175	173	174	172	173	174	175	174	171	173	174	175	174	172	171
44	177	174	175	175	172	171	172	176	172	173	172	172	173	170	170	175
43	173	171	174	168	176	172	173	173	173	174	171	174	175	173	174	174
42	175	173	171	172	170	171	176	175	178	172	174	175	175	175	175	172
41	181	179	177	172	170	170	169	179	175	174	175	174	172	175	174	175
40	188	184	179	178	176	176	176	174	172	178	172	174	173	172	174	173
39	195	191	188	186	185	183	180	177	178	175	174	176	175	174	176	176
38	200	199	197	193	190	187	185	180	176	175	180	177	175	175	176	177
37	202	202	199	202	199	194	187	180	175	179	177	176	174	175	176	173

X = *Y*	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
58	−2	−2	−2	−2	0	0	0	0	1	2	2	1	0	0	0	0
57	−2	0	0	0	0	0	0	0	1	2	2	1	0	0	0	0
56	0	0	0	0	0	0	0	1	2	2	1	0	0	0	0	0
55	0	0	0	0	0	0	0	1	2	2	1	0	0	0	0	0
54	0	0	0	0	0	0	0	1	2	1	1	0	0	0	0	0
53	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
52	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
51	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
50	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
49	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	−2	−2	0	0	0	0	0	0	1
40	0	0	0	0	1	0	0	−3	−2	−2	0	0	0	0	0	1
39	0	0	0	0	0	0	−2	−3	−3	0	0	0	0	0	1	1
38	0	0	0	0	0	−2	−3	−4	−3	−2	0	0	0	0	1	1
37	0	0	0	0	0	−3	−4	−4	−3	−2	0	0	0	0	1	1

There is nothing special about the way in which the preferences of the different methods are combined: if an obvious choice exists, it is taken, and any theory would select it. If the choice is not obvious one needs additional information, and a theory that happened to make the correct choice on marginal grounds in one image would fail in many others. The interesting point is an empirical one – that these crude selection criteria are very effective. They enable one to solve simple images completely, and almost to solve even quite difficult ones. Applying the criteria is relatively inexpensive, because the number of segments that exist at this point is much less than the number of items in the raw primal sketch. This type of filter analysis has the added attraction of being readily extendable, because the addition of extra filtering criteria simply leads to the rejection of more of the candidates at a given node.

All of the filtering criteria described above are local in the computational sense that they do not depend on the results of subsequent higher-level processes. But this does not mean that the criteria are spatially local. For example, which of the two segments 2 or 3 should be joined with segment 1 in figure 12*e*? No preference exists on purely local grounds, but a decided preference arises from the closure property of the whole figure. Only a limited degree of sensitivity to connectedness appears to be present in human visual systems (Minsky & Papert 1969, p. 73), but it is not hard to devise a detection scheme that would operate sufficiently well to help in decoding many images, while failing to provide a complete sensitivity to connectedness.

A detailed account of the selection criteria that appear to be useful will be given in a separate article, but if these methods are taken as a theory of part of our own visual systems, there is one consequence that would follow from even the sketchy account given here. If it were true that most of the time, decisions about local groupings are taken using criteria computed at roughly the same stage of the analysis, rather than by extensive use of downward-flowing information, it should be possible to find images in which a particular grouping is greatly to be preferred on most of the criteria described here, but which is nevertheless incorrect. Furthermore, if low-level decisions are indeed irrevocable (as the principle of least commitment asserts), their failure should cause severe damage to the perceptual analysis of an image. Occasionally, one finds a photograph in which the accidental alignment of contours causes this to happen, and figure 14, plate 2 shows an image whose original is misinterpreted the first time by about two people in three. The accidental alignment of the forefinger with the nose appears to be responsible for the failure. It is interesting that one does not make the same mistake the second time one views the picture; and that in the real world where stereo disparity and motion information are also available, one almost never fails at the same low level.

Transmission of unresolved nodes

The next important consequence of the principle of least commitment is that if no clear leader emerges from the group of contending possibilities, all possibilities that were not rejected are accepted. No arbitrary choices are made this early in the analysis. Nodes at which an ambiguity exists are marked, and themselves form part of the information that is sent to the next stage in the processing. The reason for doing this is that subsequent processes then have access to whatever trouble-spots exist lower down. In the image PLANT, part of the nearer leaf happens to have the same intensity as its background. Table 1*a* shows the actual intensity values in the rectangle (34, 37) to (49, 58), and table 1*b* shows the approximate edge-mask convolution values there. Although some intensity changes do exist above this area (near (44, 58)), they are insufficiently distinguished to allow the grouping methods described above to separate the two leaves. Accordingly, all of the segments are included in one form, shown together with the segments it contains in figure 13*e*. (It has been separated manually from the stem, for clarity.)

If the nodes that support this figure are maintained and can be influenced by subsequent processes, the amount of information needed to separate the two leaves is very small. For example, one decision can suffice; if it is asserted that segment (1) does not match segment (2), this information is sufficient to allow the aggregation filter network to decompose the image into the two parts shown in figures 13*f* and 13*g*. So although some downward-flowing information is needed here, the amount required is small provided that it is applied so as to use the partial results obtained at the lower level.

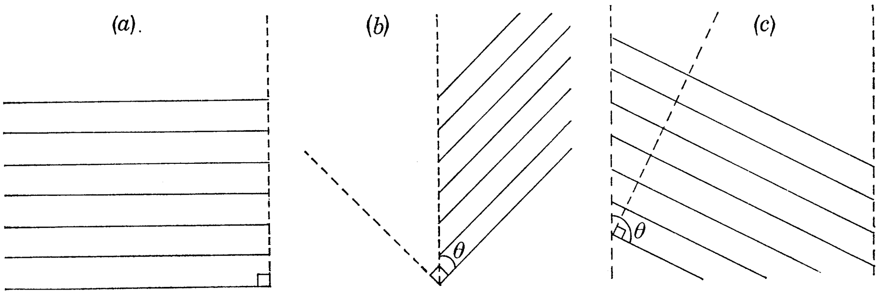


FIGURE 15. The measure of the overlap of two adjacent, parallel lines depends on an external angle, θ. In (*a*), θ is 90°, which is the value at which iteration begins in the routines that decode this type of grouping. (*b*) and (*c*) show two other values of θ.

Theta-aggregation

The techniques described above group items that possess an intrinsic orientation (or acquire one early in the processing), in a direction that approximates that local orientation. Theta-aggregation is the name we have given to the process of grouping a set of similarly oriented items in a direction that differs from their intrinsic orientation, but in a manner which uses it (e.g. figure 11*a*). The technique is to use very local grouping measures to form place-tokens that have an orientation associated with the group rather than with the local elements, and then to apply curvilinear aggregation to these tokens. The difficult part about it is that measures of the 'overlap' of two neighbouring oriented items depend upon the angle, θ, that the aggregate makes with each local unit (see figure 15). So theta determines the aggregation process, but also depends upon it. For good data, it may be quite unnecessary to know theta; aggregation of the place-tokens that each individual element defines will suffice to compute the aggregate. In general however, one will need to take into account the relation between the overall direction of the aggregate and the orientation of the local elements. Viewed from a very abstract level, this computation may be regarded as a process of solving a large number of rather simple equations.

Grouping into neighbourhoods and regions

The second category of grouping operations concerns the selection of a region by the presence there of some distinguishing local property. We first examine the nature of the local properties on which such grouping operations are based, and secondly we make some brief comments about the grouping techniques that operate on them.

Semi-local measures. From an abstract point of view, the primal sketch is simply a large body of data. There is therefore no difficulty in extracting from it certain measures and statistics, computed from the parameters that are bound to the elements of the sketch. Such measures provide a useful coarse description of a neighbourhood in the image. They can be used to control the type and depth of the analysis that is applied to a region, or to select neighbourhoods for subsequent grouping into regions. In particular, we shall assume that over moderately sized regions (0.5°–1.0° at foveal resolution) of the primal sketch the following distributions are available to processes that are capable of asking certain straightforward statistical questions of them:

D0. The total amount of contour, and number of blobs, at different contrasts and intensities.

D1. Orientation: the total number of elements at each orientation, and the total contour length at each orientation.

D2. size: distribution of the size parameters defined in the primal sketch.

D3. contrast: distribution of the contrast of items in the primal sketch.

D4. spatial density: spatial density of place-tokens defined in the different possible ways, measured by using a small selection of neighbourhood sizes.

The straightforward statistical questions referred to above include such matters as whether the distribution is uniform, or has one, two, three or more peaks; if peaks exist, where they are and their relative sizes. If the distributions are very scattered (like orientation distributions), the corresponding questions are whether the orientations are grouped in a significant way, or are roughly uniformly spread out. It has been our experience that straightforward histogram-based selection techniques suffice to drive the initial examination of an image. For example, to examine the characteristics of the orientation distribution in an image, one forms an orientation histogram based on ten degree wide orientation buckets. The figure of 10° was obtained empirically, and appears to be suitable for all images. For spatial grouping on the other hand, the scale at which one applies histogram-based techniques depends upon the place-token density of the particular image being analysed. Once again, we have not found it desirable to use elaborate statistical tests. If a property is significant, any reasonable test would detect it. If a property is marginal, no statistical model can alter the fact.

The final facility that we require is the ability to select from the image those areas or items that give rise to obvious features of these distributions. For example, in figure 18 items at an orientation of 60° are strongly predominant. We assume that items at about this orientation can be selected from the primal sketch for examination by processes that specialize in grouping such collections together. In another image, one might wish to examine first all those items whose contrast was greater than a certain value. These facilities are used only when tests indicate that they should be, and they can help the analysis of an image by greatly restricting the number of elements in the primal sketch that need to be considered by a particular process.

Boundary of a group of place-tokens. The distributions D0–D4, and the density of place-tokens obtained from items in the primal sketch, can lead to the splitting of an image into regions. The centres of figures 10*a* and 10*b* provide simple examples of this (see also Julesz 1971, pp. 105ff.). O'Callaghan (1974*a*, *b*) surveyed the literature on dot-grouping studies, and defined a local operator for obtaining boundary lines of clusters of dots. The idea is that the shape and extent of the clusters are subsequently computed from the local boundary elements.

Our experience has been that purely local methods can usually be improved by adding to them a sensitivity to the 'overall' direction of a boundary. The interaction between local and global information resembles that shown in figure 12. The overall direction of a group of place-tokens can be obtained cheaply by finding peaks in their spatial density or its gradient. Such a mechanism allows one to obtain an overall description of the shape or orientation of a group of places *before* a precise assignment of local boundary points has been made. This is relatively easy to implement, and it has the advantages of speed and economy that lead one to expect it in our own visual systems.

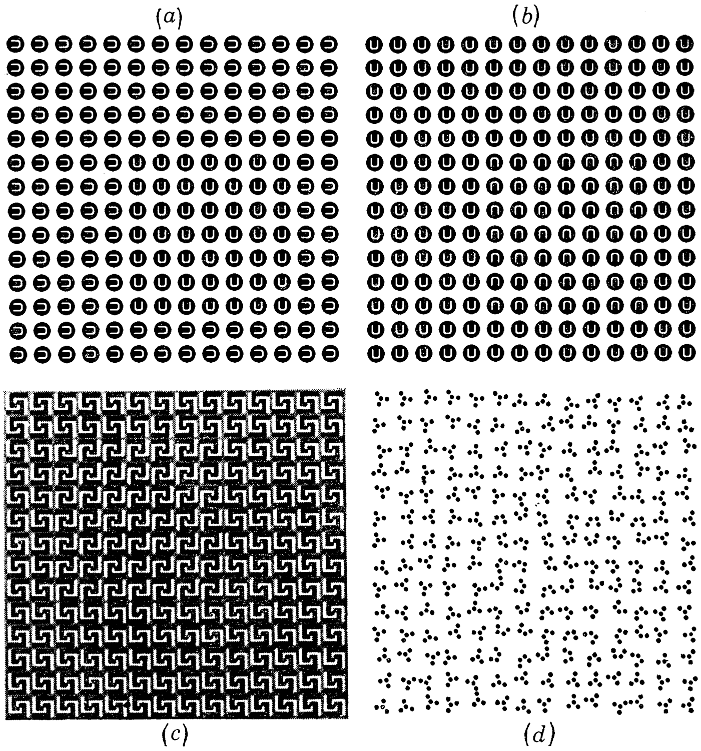


FIGURE 16. Examples of textures devised by Julesz. All four contain a square region that differs from the background. (*a*) and (*b*) obey Julesz's conjecture; in (*c*), the second order statistical structure of the square differs from that of the background, yet we cannot distinguish the two. In (*d*), the second order structure is uniform, yet we can faintly distinguish the square region. The present theory accounts for these examples, and defines a set of discriminations that neither contains nor is contained by the set of all second order discriminations.

Relation to texture-vision discrimination

There are several current ideas on texture processing. Some authors have used Fourier techniques, and in certain circumstances the spatial power spectrum can successfully separate different regions (Bacjsy 1972). Others have constructed specialized operators which sometimes discriminate between regions with different texture. Probably the earliest example of this was the Roberts gradient (Roberts 1963). The most interesting and comprehensive proposal is due to Julesz (Julesz 1962, 1975; Julesz, Frisch, Gilbert & Shepp 1973), who showed that visual textures that differ only in their third or higher order statistical structure are rarely perceptually discriminable; whereas visual textures that differ in their first or second order statistics can usually be distinguished. The important point about this finding lies in its demonstration of the essential simplicity of texture discriminations. Although it gives little insight into how the processing is implemented, it does imply that in its Volterra series expansion, all coefficients of terms whose *p*-order (Poggio & Reichardt 1976) is higher than 2 are zero.

The present theory includes texture discrimination with the other techniques for extracting forms from the primal sketch, and asserts that texture discriminations are actually implemented by the family of first order discriminations and grouping processes that act upon the primal sketch. The class of computations that these processes define differs from but overlaps considerably with the class of all second order operations operating on the original intensity array. Julesz (1975, p. 43) mentioned in an aside the possibility that texture vision may rest on 'first-order statistics of various simple feature extractors', but this idea requires the concepts of the primal sketch and of recursively applied grouping before it can be brought to fruition. The principle difference between the two approaches is that the present theory is process-oriented, since it rests on the belief that early processing of visual information is in fact implemented in this way. The second order discrimination theory provides a phenomenological description. As with many other problems of biological information processing, it will be interesting to see whether the phenomenology can be described accurately without explicitly defining the underlying computational processes.

So that the reader may form an intuitive grasp of the way in which the present theory accounts for texture vision discriminations, let us re-examine some of the textures devised by Julesz, and follow this with some examples of the texture analysis run on some natural images. First, consider figure 16. Julesz notes that in figure 16*a*, the two regions have distinct second order statistics, but not in figure 16*b*. Hence, according to his rule, the two regions are distinguishable in 16*a*, but not in 16*b*. The present theory explains this as follows: orientation measures are the only distinguishing feature of the primal sketch representation, because everything else has carefully been held constant. In figure 16*b*, the two basic elements are related by a 180° rotation, and so the orientation statistics to which they give rise are identical. Hence the two regions are indistinguishable. In figure 16*a* however, there is more contour at 0° than at 90° in the central patch, but the opposite is true in the surround. Hence the two regions are immediately distinguished.

The second example appears as figure 16*c*. Some of the modules in the pattern have been reflected about a vertical line through their centres. Their second order statistics are therefore different. This is an example in which Julesz's generalization fails. The orientation statistics of the contours, and of the local groups they form, are however unchanged, because only vertical and horizontal orientations are involved. Hence the present theory predicts that the two regions should be indistinguishable without scrutiny, as indeed they are. This establishes that the class of second order discriminations includes some operations that are not included in the class defined here.

The aggregation technique that was illustrated in figure 12 provides an example of a technique whose complexity is higher than second order. Discrimination of the distinguished region can just be made in figure 16*d*, and the reason seems to be that the dots 'string together' better there than in the background. This would be an unusual use of the aggregation techniques, but it does allow us to distinguish the region from its surround even though the second order statistical structures of the two are identical. It does not however allow us to be confident of the exact boundaries.

Examples of the analysis of some real images

In order to illustrate the usefulness of the theory, we shall now examine the results of applying it to some images. Figure 17*a* shows the primal sketch of the chair whose image appeared as figure 4*a*. The first thing to realize about this image is that it is textured at all. The texture is so simple that one easily overlooks it, yet it exists in exactly the sense of this article. The presence of the texture is suggested by the existence of three clear peaks in the orientation histogram, and the texture itself is decoded by grouping nearby items with similar orientations. Figure 17*b* and *c* shows typical results of running this procedure on this image.

Each of these aggregates can now be described simply by position, orientation and extent, and this produces a skeleton of the outline of the chair (figure 17*d*). By considering separately the structure of just one aggregate, one could go on to compute a description of the surface structure of the material out of which the chair is made. Using one autonomous technique, we have separated (but not of course solved) the problem of divining the overall three dimensional shape of the chair from the analysis of its surface properties. This ability is vital if the organization of subsequent analysis is to be modular.

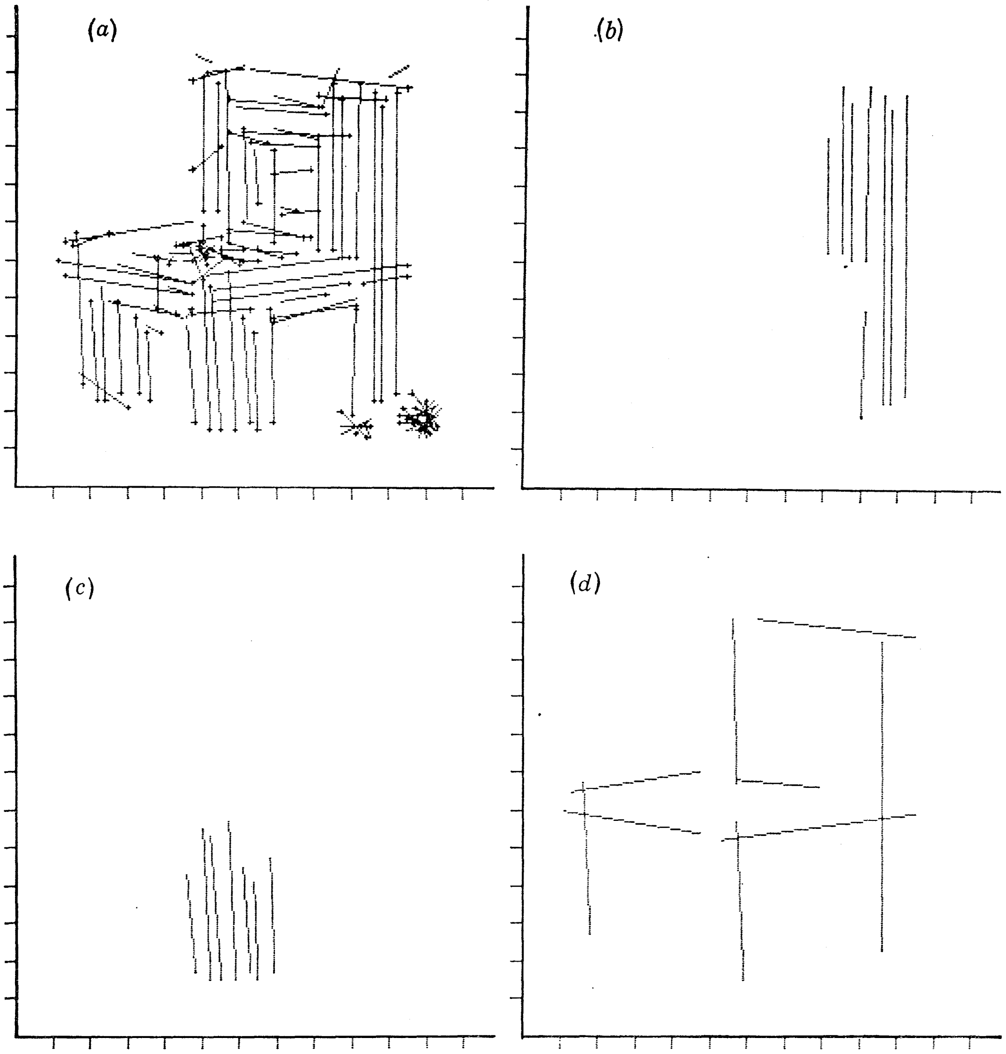


FIGURE 17. The spatial information of the primal sketch of the image CHAIR (figure 4*a*, plate 1) is shown in (*a*). (*b*) and (*c*) show two units that emerge after aggregation, and (*d*) gives the skeleton of the chair to which this aggregation leads. (This skeleton was obtained by selecting the longest edge from each aggregate, and adding the edge whose centre lies at (30, 67).) By using the texture that is present in the image, the problem of divining the three dimensional shape of the object has been separated from the problem of recognizing its surface structure (one takes (*d*) as its data, the other takes units like (*b*) or (*c*)). No downward-flowing information was necessary to accomplish this.

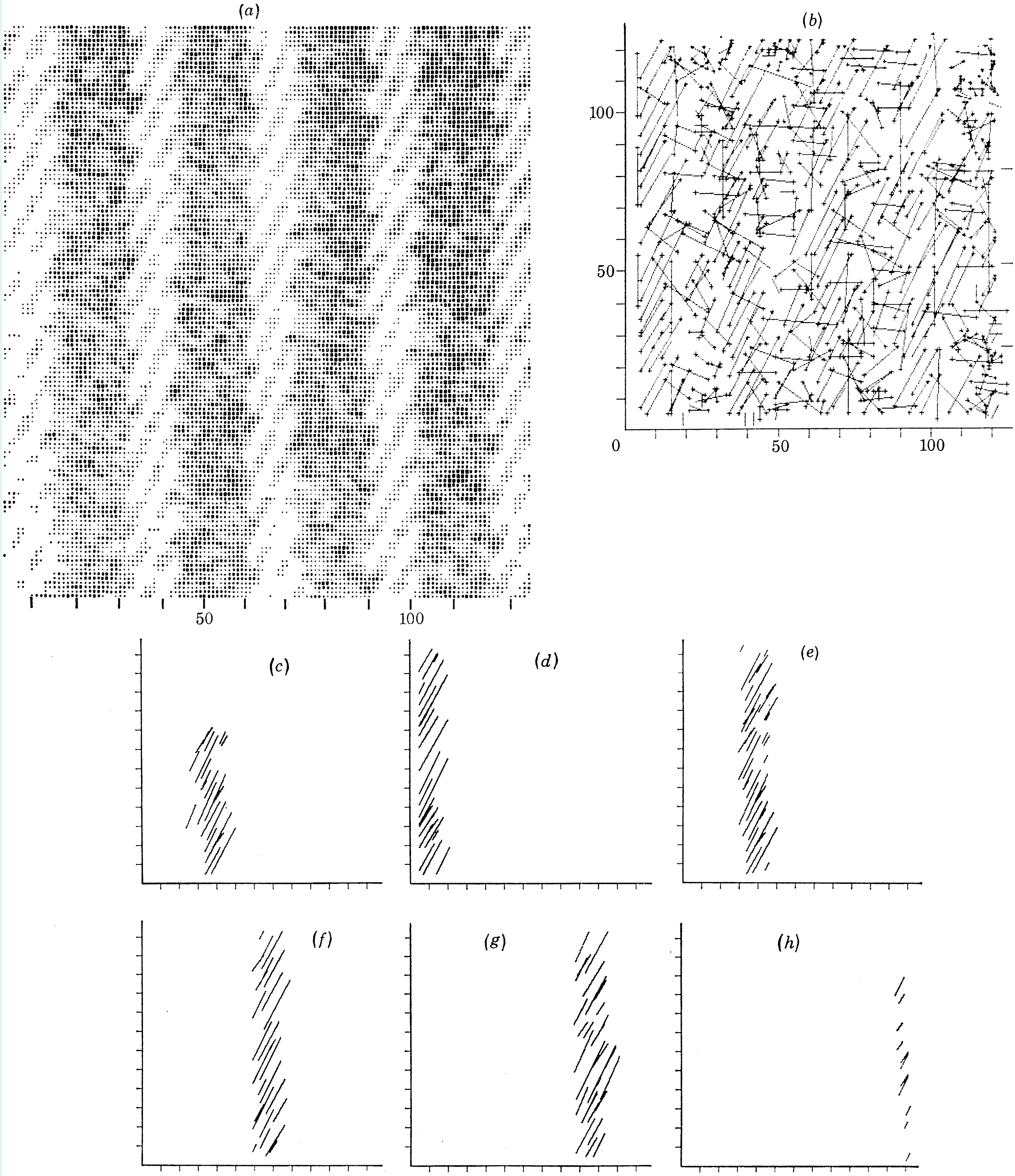


FIGURE 18. The image shown in figure 4*d*, plate 1 (taken from Brodatz (1966), plate D11), has been printed in (*a*). (*b*) shows a rendering of the spatial component of the primal sketch. The predominance of items at an orientation of 60° (see table 2) causes θ-aggregation to be attempted at this orientation. Initial grouping produces the aggregate (*c*). From this, θ is found, and the aggregation process then extracts the stripes successfully, as shown in (*d*)–(*h*).

The next example shows a difficult case of theta-aggregation. The image (figure 4*d*) is not very contrasty because it was taken from a photograph (Brodatz 1966, plate D11). The intensity values have been printed in figure 18*a*, and figure 18*b* shows the spatial component of the primal sketch. Contours of all intensities lengths, and orientations are shown, and as one would expect from an image of this complexity, 18*b* has a somewhat messy appearance. Part of the mess can be removed by excising elements responsible for the lowest-contrast peak in the contrast-

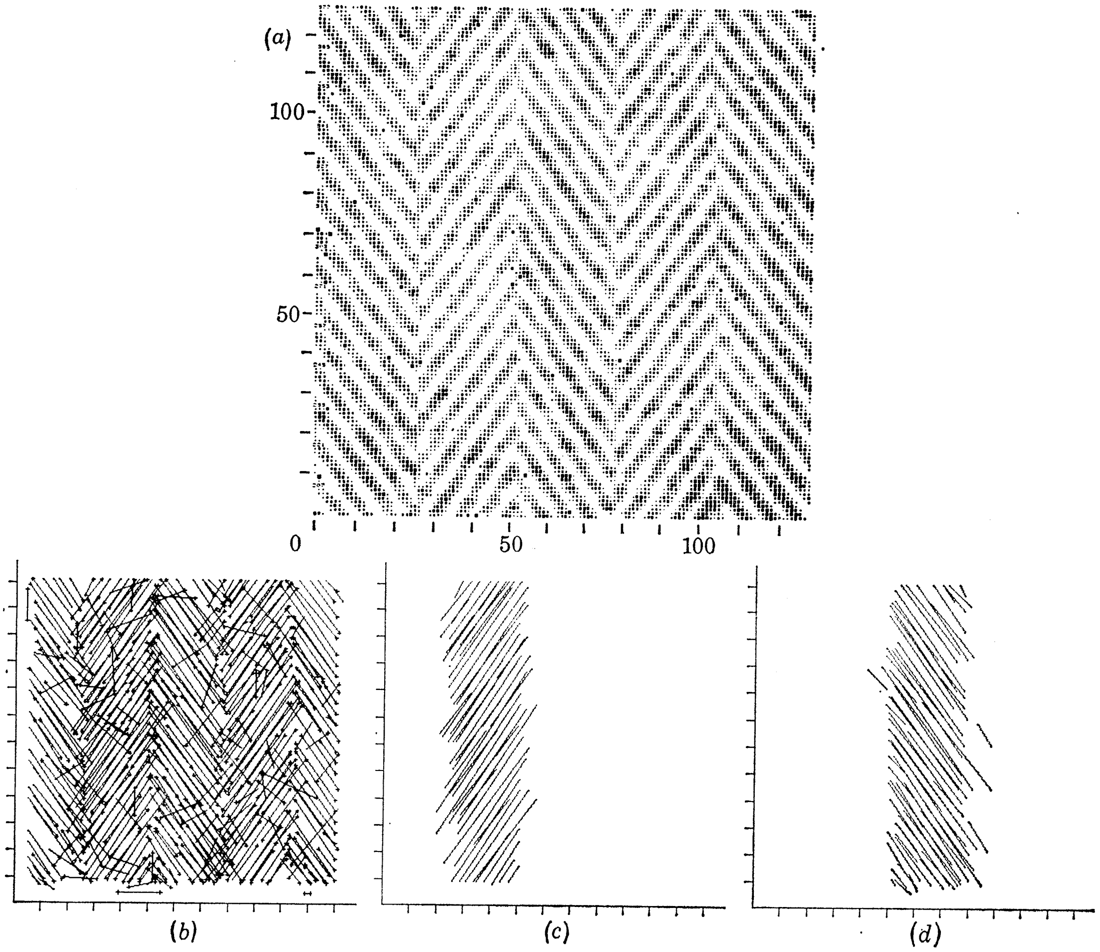


Figure 19. The analysis of this herring-bone pattern (figure 4*e*, plate 1) demonstrates that the methods for distinguishing two texture regions do not depend on their having different average reflectances. (*a*) shows the printed image, and (*b*) the spatial component of the primal sketch. Typical extracted stripes are shown in (*c*) and (*d*).

Table 2. First order measures taken over the primal sketch can control the execution of grouping techniques

(This table shows the orientation statistics of the primal sketch shown in figure 18. For the purpose of illustration, the orientations have been divided into disjoint buckets 15° wide, and the total amount of contour and number of primal sketch elements are shown for each of these buckets. Any criterion would judge 60° to be an important orientation. The processor therefore tries to group contours having this orientation.)

orientation deg	0	15	30	45	60	75	90	105	120	135	150	165
number of items	64	7	14	16	161	27	42	15	25	28	34	16
total contour length	632	64	132	116	2213	186	600	118	198	304	331	138

distribution histogram, but the crucial clues come from the orientation distribution. Table 2 provides rough information about the amount of contour that is present at each orientation, from which it is evident that items at an orientation of around 60° predominate. The average length of items at this orientation is 13. These coarse measures cause the texture analyser to attempt to group the edges at this orientation. Initially, the direction along which grouping should take place is unknown, so stringent local grouping parameters are used. This leads to the primary cluster shown in figure 18*c*. From this, an overall direction is obtained (−88°), and curvilinear aggregation then groups the items into the stripes shown in 18*d–h*. This completes primary texture processing. Once the primary stripes have been obtained, the same analysis operating recursively on tokens for these stripes serves to relate them to one another. Notice that in this particular image, some of the stripe information has been picked up directly from the intensity values (see figure 18*b*). This would not be true of a more herring-bone texture, and the analysis does not depend upon it. Our present system is successful at processing herring-bone textures of similar complexity in which the two types of stripe have the same average reflectance. Figure 19 demonstrates this. It shows the analysis of figure 4*e*, plate 1 which is a fragment of Dr Eric Sandewall's waistcoat.

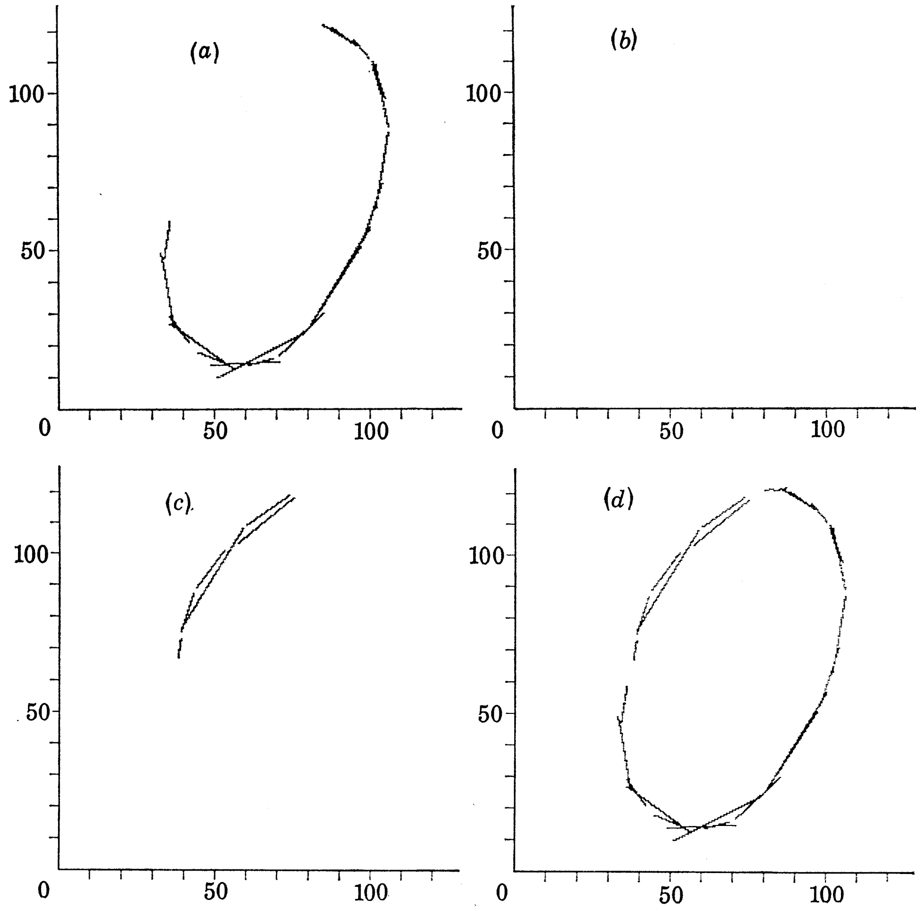


FIGURE 20. The first two stages of curvilinear aggregation have been run on the primal sketch of the rod shown in figure 7, and they produced the elements (*a*), (*b*) and (*c*). Once larger units have been obtained, the governing parameters can be relaxed, and the elliptical form (*d*) is obtained by the third step. Up to this point, the system has neither computed nor used any descriptor of the form's overall shape.

Finally, I give two examples of images that are simple enough for the aggregation techniques to extract the important forms unaided. The local elements of the primal sketch of the rod of figure 6 are grouped by the first two stages of curvilinear aggregation into the units shown in figure 20 *a–c*. The third stage assembles them into the form shown in 20 *d*. The reason why the first two stages cannot complete the job is because of the alternatives near (33, 60), and because the contrast across the top-left portion of the form has the opposite sign from the contrast elsewhere.

Several types of analysis have been applied to the image of a toy bear (figure 21). The half-tone image (figure 4*f*) has been printed in 21 *a*, and the intensity map is given in 21 *b*. The

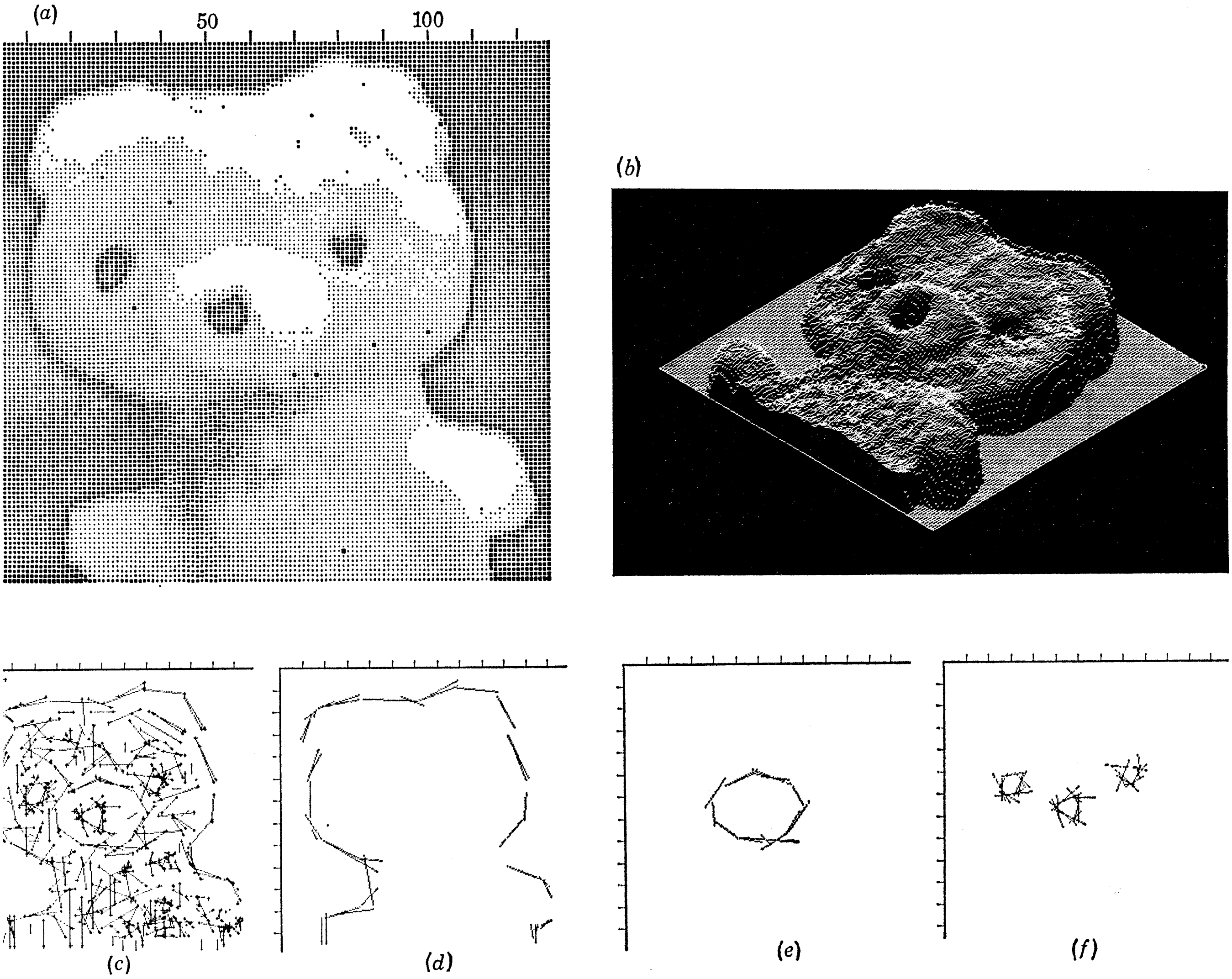


FIGURE 21. The image of a toy bear (figure 4*f*, plate 1) has been printed in (*a*), and its intensity map appears in (*b*). The spatial component of the primal sketch is illustrated in (*c*). The three principal forms extracted from (*c*) appear in (*d*), (*e*) and (*f*). The items in (*f*) are classed as BLOBS, and the configuration that they form is recognized as a VEE (figure 11 *h*) with modifier FLAT. The axis relative to which this configuration was computed is the vertical (default value). The outline of the bear (*d*), and of his muzzle (*e*) are simple enough to have been extracted using only the techniques described in this article. The closed form property was used to help decide between competing segments at coordinate (80, 65). (The vertical appears as the negative x axis because this image was taken with the camera on its side.)

primal sketch of this image is represented by 21*c*. The blobs extracted from this image appear in figure 21*f*, and the routines for describing the spatial disposition of a small number of places recognize that these form a (VEE FLAT) configuration (cf. figure 11*h*), described relative to the default vertical axis. The contours that form the bear's face appear in 21*d*, and 21*e* shows his muzzle. The extraction of the muzzle made use of the closed form property, as well as discrepancies in contrast and fuzziness, while choosing between rival segments near coordinate (80, 65).

Discussion

Perhaps the most novel aspect of these ideas is the notion that the primal sketch exists as a distinct and circumscribed symbolic entity, computed autonomously from the image, and operated on repeatedly by a number of local geometrical processes, semi-local measures, and first order discriminations. The underlying reason why one needs to compute such a thing is that in some sense a description like the primal sketch is much closer to what is really there (i.e. changes in reflectance) than the values of edge-shaped or bar-shaped mask convolutions, which form a large and confusing set of primary measurements. It would be almost impossible to deal with so huge a mass of data unless it were first organized into a readable format.

The storage into which the primal sketch is written is the direct analogue for the class of images studied here of the Cyclopean retina that Julesz (1971) wrote of for binocular vision. More subjectively, what it holds corresponds very closely to the 'image' that one is conscious of. This reflects the computational hypothesis that all subsequent analysis reads the primal sketch, not the data from which it was computed. The primal sketch therefore acts in a genuine sense as the interface at which visual analysis becomes a purely symbolic affair.

Implications for neurophysiology

The images studied here are impoverished by their inherent lack of movement or binocular disparity. Extreme caution is needed when attempting to make predictions from such a theory, because of the power of these two types of information. For example, a linear cell with a centre-surround receptive field is a hopeless blob-asserter on its own. The frog's fly-catching system only works because the additional constraint of relative motion is added (Barlow 1953; Lettvin, Maturana, McCulloch & Pitts 1959). Movement information together with some extra circuitry might even turn a linear simple cell with a bar-shaped receptive field into a passable detector of bars in an image. But a simplistic scheme of this sort, though possibly acceptable to a cat, would be of little use for deciphering a motionless scene. It is therefore reasonable to expect that something like a primal sketch is computed, at least by the higher primates. If it is, the cells that represent the primal sketch should exhibit the consequences of algorithms like peak-matching, the selection criterion, and the (otherwise surprising) inter-orientation interactions that are central to its construction. One would also expect grouping processes that use disparity or motion information to take as their input the primal sketch and at least some of the classes of tokens obtained from it (Marr 1974).

At a higher level, one would expect to find experimental evidence of the aggregation processes that the theory predicts should act upon the primal sketch to decompose it into unit forms. Some of these processes have natural neural representations, and some do not. For curvilinear and theta-aggregation, one would expect to find a cell that marks the overall direction of aggregation independently of the orientation of the local elements. One would also expect to find cells that represent place-tokens (recognizable by their insensitivity to what is at the place); and cells for carrying pieces of the local first order and spatial-density measures that are important for texture-based definition of regions. The design of the most likely neural representation of these processes is not straightforward.

The influence of higher-level knowledge and of purpose on visual information processing

There are two broader implications of the theory that are worth mentioning. First, the four principles stated at the beginning of the article have survived intact, and their guidance has been valuable. The principle of least commitment has played an especially important rôle, by its pressure on us to design a system that does not usually do anything wrong. It caused us to abandon ideas about 'trigger features' in favour of the computation of a 'true' description, which led in turn to the gradual elucidation of the processes that are necessary to read it. The result is bulky rather than complex, and requires prodigious computing power but little computing sophistication (it could be implemented without difficulty in a stackless machine). There can however be no doubt that in terms of sheer processing power, the human visual system must be spectacularly well-endowed.

The second implication of interest concerns the structure of subsequent recognition processes. If non-attentive vision may be implemented successfully by approximately the set of methods defined in this article, it means that visual 'forms' can usually be extracted from the image by using knowlege-free techniques. In other words, the extraction of a visual form can usually *precede* its description. From this it follows that it is usually easy to compute a *coarse description* of a form before having any idea about what the form is.

If this is true, it greatly simplifies the design of subsequent recognition processes, because it means that they too can be made modular. For example, the ability to compute a coarse description of a form allows one to describe the shape of a forest without first computing detailed descriptions of all the trees; or to compute the shape of the cluster of blobs that forms a distant village independently of deciding that some of those blobs are actually buildings and that the cluster is therefore a village. In the more mundane example of figure 21, one can compute that the overall shape of the top form is roughly ovoidal without first having to segment out and describe separately the bumps that are the bear's ears. The autonomy of early visual processing permits the rôle of higher level knowledge to be very restricted, and different in kind from its intervention in programs like Shirai's (1973). Downward-flowing information will not affect the line-finding stage (the computation of the primal sketch) at all. Its most usual *modus operandi* is in choosing which processes are to be used to read the primal sketch – for example by specifying which texture predicate should be used on the image to select the parts of current interest. It can also apply certain limited kinds of flags to critical segments during their aggregation into forms (as in the image PLANT). The coupling between higher-level knowledge and the form-extraction processes is however much weaker than the coupling between the different form-extraction processes.

It is clearly desirable to have some control over which of the possible forms in a figure should be delivered at a given moment from the primal sketch. For example, in the image BEAR there are three possible major forms; the outline of the head, the muzzle, and the three blobs that represent his eyes and nose. It seems probable that only one of these should be made available at a time, and this in turn raises interesting questions about the order in which it is done, the way in which the three forms and their relative positions are described, and the way in which those descriptions trigger a larger datastructure and are absorbed by it. In living systems, which are powerful enough to operate in real time, the control of the direction of gaze may be rather closely related to the order in which these events take place.

This study would not have been possible without the advanced and flexible computing facilities that are available at the Artificial Intelligence Laboratory. I thank Professor M. Minsky and Professor S. Papert for inviting me to the laboratory; Hassan Alam, Gary Dudley, and especially Ken Forbus for programming assistance; Dr B. Julesz and Pion Publications for permission to reproduce figure 16; Bruce Schatz for devising a wide range of patterns for texture vision discriminations; and Karen Prendergast for preparing the drawings. The material described in this paper covers M.I.T. A.I. Lab. Memos 324 and 334, and a preliminary report has appeared (Marr 1976*b*). I thank Cold Spring Harbor Laboratory for permission to reproduce figures 8, 18 and 20 from there. This work was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense, and monitored by the Office of Naval Research under Contract number N00014-75-C-0643.

References

Bajcsy, R. 1972 Computer identification of textured visual scenes. *Stanford A.I. Lab. Memo.* 180.

Barlow, H. B. 1953 Summation and inhibition in the frog's retina. *J. Physiol., Lond.* **119**, 56–68.

Brodatz, P. 1966 *Textures: a photographic album for artists and designers.* New York: Dover Publications.

Freuder, E. C. 1974 A computer vision system for visual recognition using active knowledge *M.I.T.A.I. Lab. Technical Report* 345.

Harmon, L. D. & Julesz, B. 1973 Masking in visual recognition: effects of two-dimensional filtered noise. *Science N. Y.* **180**, 1194–1197.

Herskovits, A. & Binford, T. O. 1970 On boundary detection. *M.I.T.A.I. Lab. Memo* 183.

Horn, B. K. P. 1973 The Binford–Horn LINEFINDER. *M.I.T.A.I. Lab. Memo.* 285.

Hubel, D. H. & Wiesel, T. N. 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol., Lond.* **160**, 106–154.

Hueckel, M. H. 1971 An operator which locates edges in digitized pictures. *J. Ass. Comput. Mach.* **18**, 113–125.

Hueckel, M. H. 1973 An operator which recognizes edges and lines. *J. Ass. Comput. Mach.* **20**, 634–647.

Julesz, B. 1962 Visual pattern discrimination. *IRE Transactions of Information Theory,* **IT-8**, 84–92.

Julesz, B. 1971 *Foundations of cyclopean perception.* Chicago: The University of Chicago Press.

Julesz, B. 1975 Experiments in the visual perception of texture. *Sci. Am.* **232**, 34–43 (April issue).

Julesz, B., Frisch, H. L., Gilbert, E. N. & Shepp, L. A 1973 Inability of humans to discriminate between visual textures that agree in second-order statistics–revisited. *Perception* **2**, 391–405.

Kanizsa, G. 1955 Margini quasi-percettivi in campi con stimulazioni omogenea. *Rivista di Psicologia* **49**, 7–30.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S. & Pitts, W. H. 1959 What the frog's eye tells the frog's brain. *Proc. Inst. Radio Engrs* **47**, 1940–1951.

McCarthy, J. *et al.* 1963 *LISP* 1.5 *Programmer's Manual.* Cambridge Mass.: The M.I.T. Press.

Macleod, I. D. G. 1970 On finding structure in pictures. In *Picture language machines* (ed. S. Kaneff), p. 231. New York: Academic Press.

Maffei, L. & Fiorentini, A. 1973 The visual cortex as a spatial frequency analyser. *Vision Res.* **13**, 1255–1267.

Marcus, M. P. 1974 Wait-and-see strategies for parsing natural language. *M.I.T.A.I. Lab. Working Paper* 75.

Marr, D. 1971 Simple memory: a theory for archicortex. *Phil. Trans. R. Soc. Lond.* B **262**, 23–81.

Marr, D. 1974 A note on the computation of binocular disparity in a symbolic, low-level visual processor *M.I.T.A.I. Lab. Memo* 327.

Marr, D. 1976*a* Technical problems in the early processing of visual information. (In preparation.)

Marr, D. 1976*b* Analyzing natural images: a computational theory of texture vision. *Cold Spring Harbor Symp. Quant. Biol.* **40**, 647–662.

Minsky, M. & Papert, S. 1969 *Perceptrons.* Cambridge, Mass.: M.I.T. Press.

O'Callaghan, J. F. 1974*a* Human perception of homogeneous dot patterns. *Perception* **3**, 33–45.

O'Callaghan, J. F. 1974*b* Computing the perceptual boundaries of dot patterns. *Computer graphics and image processing* **3**, 141–162.

Poggio, T. & Reichardt, W. 1976 Visual control of orientation behaviour in the fly. Part II: towards the underlying neural interactions. *Quart. Revs Biophys.* (In the press.)

Ratliff, F. 1965 *Mach bands: quantitative studies on neural networks in the retina.* San Francisco: Holden-Day.

Roberts, L. 1963 Machine perception of three-dimensional solids. *Technical Report* **315**, Lincoln Laboratory, M.I.T.

Rosenfeld, A. & Thurston, M. 1971 Edge and curve detection for visual scene analysis. *I.E.E.E. Trans. Comput.* C-20, 562–569.

Rosenfeld, A., Thurston, M. & Lee, Y. H. 1972 Edge and curve detection: further experiments. *I.E.E.E. Trans. Comput.* C-21, 677–715.

Shirai, Y. 1973 A context-sensitive line finder for recognition of polyhedra. *Artificial intelligence* **4**, 95–120.

Waltz, D. 1975 Understanding line drawings of scenes with shadows. In: *The psychology of computer vision* (Ed. P. H. Winston), pp. 19–91. New York: McGraw-Hill.

Wertheimer, M. 1923 Untersuchungen zur Lehre von der Gestalt, II. *Psychol. Forsch.* **4**, 301–350.

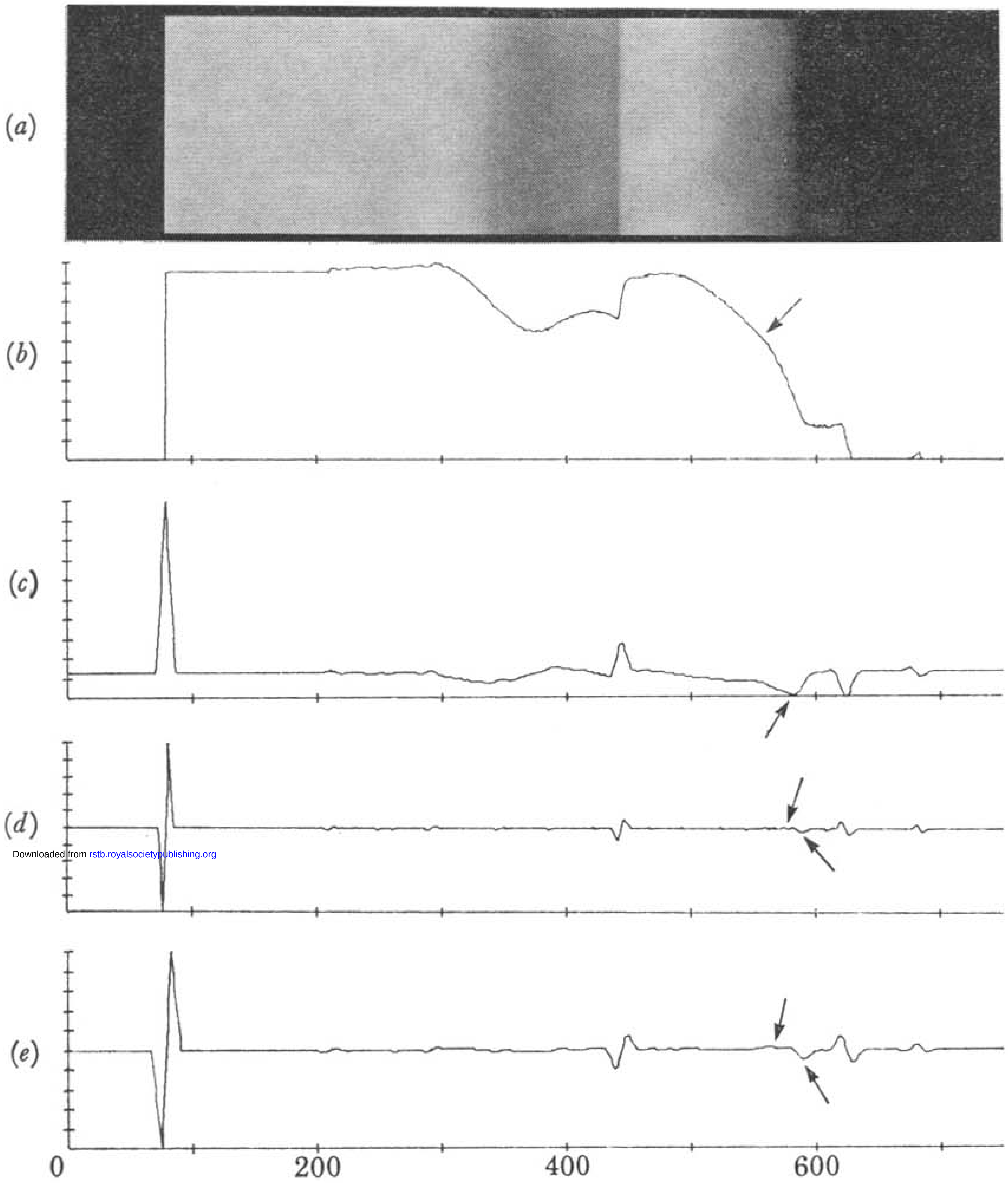
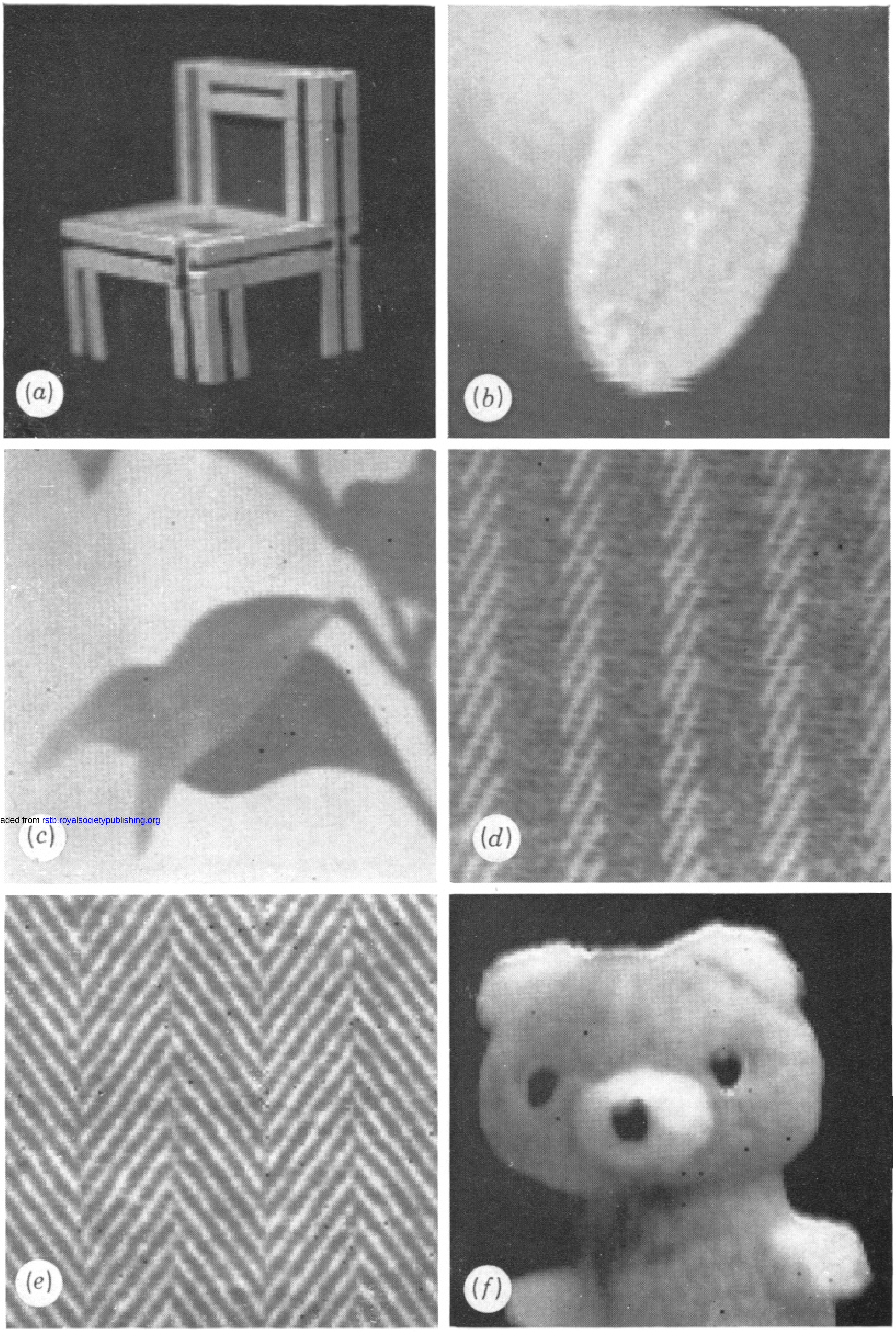


FIGURE 3. The intensity distribution exhibited in (*a*), whose profile appears in (*b*), was obtained by illuminating a curved piece of white paper from one end, and viewing it from above. Its description, computed by using an edge-mask of panel-width 8 (*c*), and bar masks-of panel-widths 4 (*d*) and 8 (*e*), is as follows:

EDGE (POSITION 80) (CONTRAST 136) (FUZZ SHARP)
EDGE (POSITION 212) (CONTRAST 3) (FUZZ 4)
EDGE (POSITION 292) (CONTRAST 2) (FUZZ SHARP)
EDGE (POSITION 435) (CONTRAST −3) (FUZZ 4)
EDGE (POSITION 444) (CONTRAST 25) (FUZZ 5)
EDGE (POSITION 464) (CONTRAST 2) (FUZZ 4)
EDGE (POSITION 490) (CONTRAST 1) (FUZZ 4)
EXTENDED-EDGE (POSITION 582) (CONTRAST −12) (FUZZ 9)
(the peaks giving rise to this edge are marked with arrows)
EDGE (POSITION 624) (CONTRAST −20) (FUZZ 6)
EDGE (POSITION 676) (CONTRAST 3) (FUZZ 4)
EDGE (POSITION 684) (CONTRST −4) (FUZZ 4)
SHADING-EDGE (POSITION 570) (CONTRAST −14) (WIDTH 67)
SHADING-EDGE (POSITION 391) (CONTRAST 4) (WIDTH 36)
SHADING-EDGE (POSITION 339) (CONTRAST −8) (WIDTH 73)



IGURE 4. This figure provides a high quality reproduction of the six images discussed in the text. (*a*) and (*b*) were taken with a considerably modified Information International Incorporated Vidissector, and the rest were taken with a Telemation TMC-2100 vidicon camera attached to a Spatial Data Systems digitizer (Camera Eye 108). The full dynamic range from black to white is represented by 256 grey-levels. The images reproduced here were created by an Optronics P150ohPhotowriter from intensity arrays that measured 128 elements square. This size of intensity array corresponds to viewing a 1 in square at 5 ft with the human retina. The image of the period at the end of this sentence probably covers more than 40 retinal receptors. The reader should view the images from a distance of about 5 ft when assessing the performance of the programs. In the interests of clarity, these intensity arrays have been displayed in two other ways (where helpful). They have been printed on a Xerographic printer using a font of 16 grey-levels; and they have been displayed as a three dimensional graph, in which the *z* coordinate represents intensity. These displays appear in the figures.

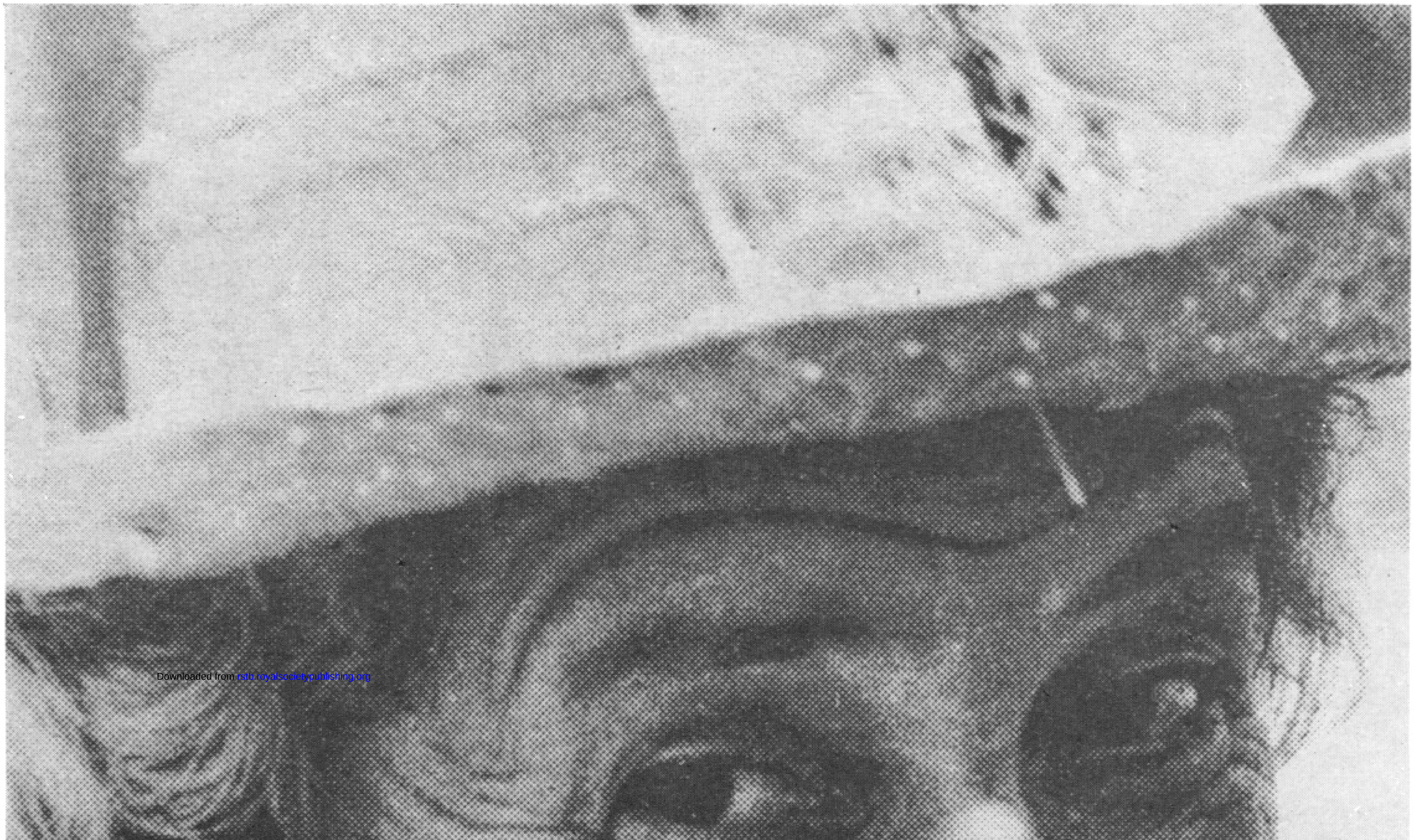


FIGURE 14. About two people in three fail to perceive the original of this image correctly the first time. The failure is caused by the accidental alignment of the subject's forefinger and nose. This failure shows that simple local processes are important during the analysis of an image, and that delivery by them of an incorrect grouping is not a normal event. This is good evidence against the hypothesis that early visual processing is designed around a failure-driven control structure. The fact that one does not make the same mistake a second time shows that some downward-flowing information can affect early processing. Only a small amount would be required to prevent recurrence of the error.

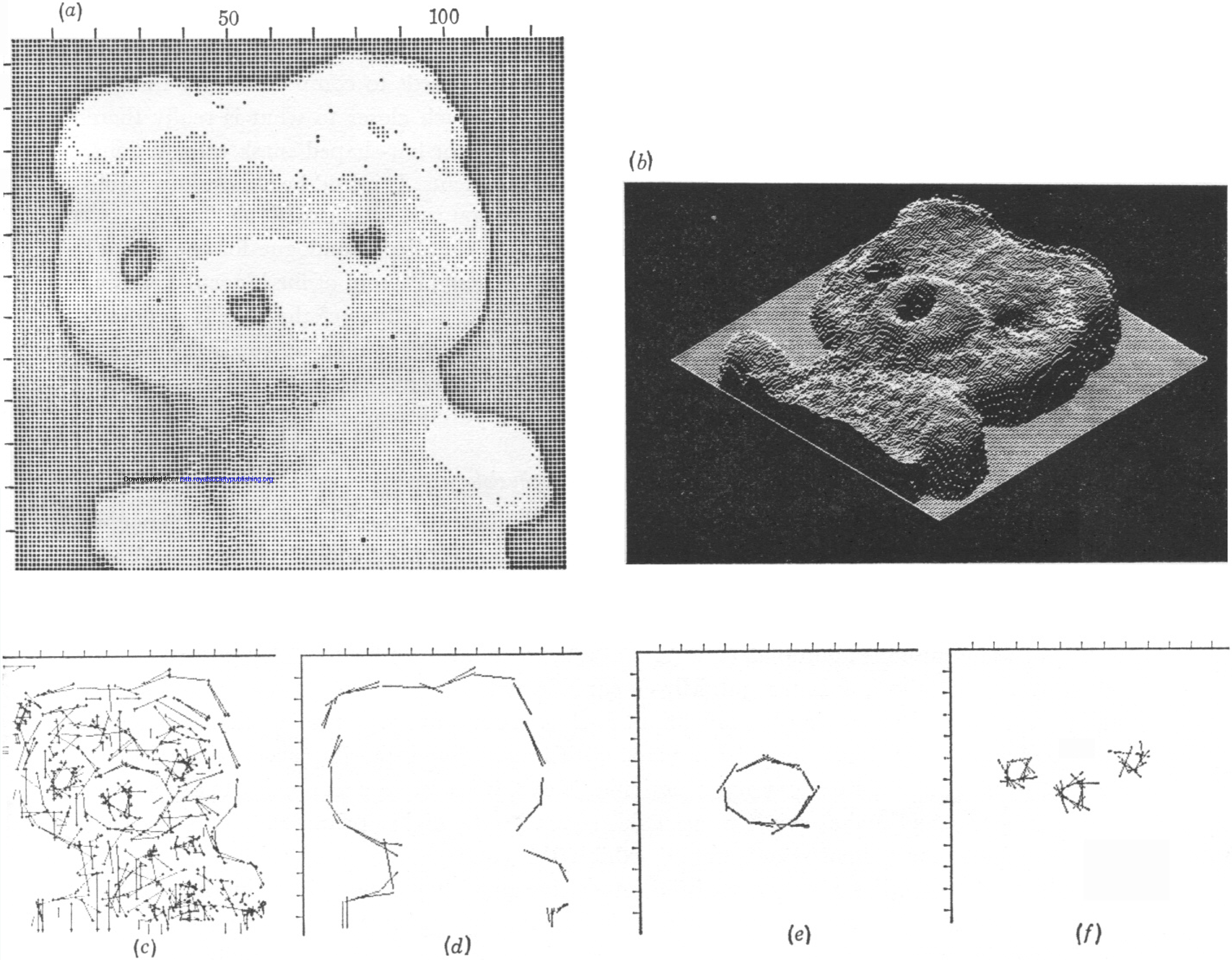


FIGURE 21. The image of a toy bear (figure 4*f*, plate 1) has been printed in (*a*), and its intensity map appears in (*b*). The spatial component of the primal sketch is illustrated in (*c*). The three principal forms extracted from (*c*) appear in (*d*), (*e*) and (*f*). The items in (*f*) are classed as BLOBS, and the configuration that they form is recognized as a VEE (figure 11 *h*) with modifier FLAT. The axis relative to which this configuration was computed is the vertical (default value). The outline of the bear (*d*), and of his muzzle (*e*) are simple enough to have been extracted using only the techniques described in this article. The closed form property was used to help decide between competing segments at coordinate (80, 65). (The vertical appears as the negative *x* axis because this image was taken with the camera on its side.)